
Stochastic timeseries analysis in electric power systems and paleo-climate data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Leonardo Rydin Gorjão

aus Lissabon, Portugal

Berichterstatter:

J-Prof. Dr. Dirk Witthaut

Prof. Dr. Joachim Krug

Prof. Dr. Giovanni Filatrella

Vorsitzender der Prüfungskommission:

Prof. Dr. Tobias Bollenbach

Tag der letzten mündlichen Prüfung: 18.02.2021

Abstract

In this thesis a data science study of elementary stochastic processes is laid, aided with the development of two numerical software programmes, applied to power-grid frequency studies and Dansgaard–Oeschger events in paleo-climate data.

Power-grid frequency is a key measure in power grid studies. It comprises the balance of power in a power grid at any instance. In this thesis an elementary Markovian Langevin-like stochastic process is employed, extending from existent literature, to show the basic elements of power-grid frequency dynamics can be modelled in such manner. Through a data science study of power-grid frequency data, it is shown that fluctuations scale in an inverse square-root relation with their size, alike any other stochastic process, confirming previous theoretical results. A simple Ornstein–Uhlenbeck is offered as a surrogate model for power-grid frequency dynamics, with a versatile input of driving deterministic functions, showing not surprisingly that driven stochastic processes with Gaussian noise do not necessarily show a Gaussian distribution.

A study of the correlations between recordings of power-grid frequency in the same power-grid system reveals they are correlated, but a theoretical understanding is yet to be developed. A super-diffusive relaxation of amplitude synchronisation is shown to exist in space in coupled power-grid systems, whereas a linear relation is evidenced for the emergence of phase synchronisation.

Two Python software packages are designed, offering the possibility to extract conditional moments for Markovian stochastic processes of any dimension, with a particular application for Markovian jump-diffusion processes for one-dimensional timeseries.

Lastly, a study of Dansgaard–Oeschger events in recordings of paleoclimate data under the purview of bivariate Markovian jump-diffusion processes is proposed, augmented by a semi-theoretical study of bivariate stochastic processes, offering an explanation for the discontinuous transitions in these events and showing the existence of deterministic couplings between the recordings of the dust concentration and a proxy for the atmospheric temperature.

Contents

1	Introduction	1
1.1	Energy systems and power-grid frequency dynamics	2
1.2	Paleo-climatic transitions and Dansgaard–Oeschger events	7
1.3	Stochastic processes	9
1.3.1	Applications in power-grid frequency studies	9
1.3.2	Applications in paleo-climate studies	10
1.4	Overview of relevant publications	10
2	Publications	15
2.1	Stochastic analysis and modelling of power-grid frequency dynamics . . .	15
2.1.1	Publication #1	15
2.1.2	Publication #2	32
2.1.3	Publication #3	44
2.2	Spatio-temporal analysis of power-grid frequency dynamics	51
2.2.1	Publication #4	51
2.2.2	Publication #5	78
2.3	Jump-diffusion processes and analysis of paleo-climatic transitions and Dansgaard–Oeschger events	90
2.3.1	Publication #6	90
2.3.2	Publication #7	103
2.3.3	Publication #8	124
3	Conclusions	137
3.1	Power systems and power-grid frequency	137
3.2	Bivariate jump-diffusion processes and Dansgaard–Oeschger events . . .	143
3.3	Software development	145

Bibliography	147
A Author contributions	A 3
B Data availability	B 7
B.1 Power-grid systems	B 7
B.1.1 List of sources	B 8
B.2 Paleo-climate high-frequency data	B 8
B.2.1 List of sources	B 8
C Erklärung zur Dissertation	C 9

Chapter 1

Introduction

Stochastic noise is ubiquitous in physical systems. Its presence embodies a collection of phenomena: external fluctuations, high-frequency couplings in the system's elements, and from a measurement perspective, instrumental noise and uncertainty. Commonly, noise is considered as a drawback as it can impede the stability or observability of a system. Yet, the stochastic characteristics carry an enormous scientific and application value as much of the processes' intrinsic characteristics manifest themselves through their noise.

This dissertation touches on modelling and analysis of continuous-time stochastic processes with applications in energy systems and paleo-climate data. It comprises a development of non-parametric estimators for continuous-time stochastic processes in N dimension, as well as the development of a non-parametric estimators for bivariate Markovian Poissonian jump-diffusion processes. The methods developed and numerically implemented are applied in power-grid frequency studies and Dansgaard–Oeschger events in recordings of paleo-climate data.

This thesis begins with a short introduction to the applications discussed in this thesis, i.e., power-grid frequency dynamics and paleoclimate data, a prologue to continuous-time stochastic processes employed, and an overview of the publications included in this thesis—either published, submitted, or in preparation—with a short abstract of each publications. The second chapter contains all scientific publications grouped into three sections. Firstly, a section on data-driven power-grid frequency modelling, where the relevant publications are included [1, 2, 3]. Secondly, a spatio-temporal study of power-grid frequency augmented with synchronous recordings [4, 5]. Thirdly, a section addressing fast paleo-climatic transitions, in specific, Dansgaard–Oeschger event in the Last Glacial

Period, and the application of bivariate jump-diffusion processes [6, 7, 8]. The last chapter concludes with a critical examination of the publications as well as an examination of contemporaneous works and their implications. A note on authorial contributions and data used and collected during the thesis is found in the appendices.

1.1 Energy systems and power-grid frequency dynamics

Energy systems, and in particular, power-grid systems are the technical backbone of modern society. The access to electricity, ubiquitous in the developed world, is central to the functioning of society, for the most basic human needs to the most advanced technological and industrial applications, rely on the access to electric power [9]. Underlying what for many has become a commodity are complex control mechanisms ensuring that electricity is robustly available to everyone [10]. These systems, power-grid systems, are amongst the most complicated human-made structures, sometimes spanning entire continents and traditionally operate under strict control mechanisms [11].

Maybe the most remarkable feature of the dynamics of power grids is the emergence of self-organised synchronisation on vast spatial scales up to thousands of kilometres. All inertial generators across a power grid rotate with the same frequency (or integer multiples thereof), at 50 Hz in Europe [12]. One curious aspect here is that unlike many other commodity networks, like gas or water supply, power cannot be easily stored. Although at a small scale—to operate mobile phones or run a laptop—batteries or other power storage exist, at the scale of countries or continents, power is generated and consumed simultaneously. This fact makes these systems unique, as they have to ensure that the customers' desires are met at each instant, i.e., that roughly at any chosen minute of the day, the power being produced within a power grid is simultaneous being consumed. Thus, a large market structure exists behind these systems, ensuring that each producer—be it a large nuclear power plant, a hydro generator, or a collection of wind turbines—can sell power to the ever-present and ever-changing consumers [13].

Noticeably, the power-grid frequency carries a mark of each of these elements. As mentioned, in a power-grid system, a nominal angular rotation of synchronous generators must be ensured [14, 15, 16]. At any moment one expects to be able to withdraw energy from a power grid—on a local power socket or over a larger power cable—with a fixed frequency: fifty cycles per second. Now as this frequency is ensured to be at the desired

nominal frequency by a collection of coupled rotating masses at each power plant, if there is a lack of power being produced, the rotation of these masses starts slowing down [10]. Likewise, an excess of power production accelerates the rotation of these generators. These changes of power generation are proportional to the frequency deviations and these deviations can easily be seen by examining power-grid frequency recordings—one of the central timeseries analysed in this thesis.

Naturally, to ensure that the power-grid frequency is kept at the desired nominal value a set of control mechanisms are in place [17]. In some sense, the first one was already mentioned: Large coupled rotating masses serve as inertia in these systems, ensuring that the rate at which frequency deviation grows is bounded, thus leaving time for the remaining control mechanisms to be activated. [18]. Alongside the inertia in the system, a subset of power producers, mainly fast-reaction gas fired power plants aided by battery storage, are kept in reserve to add or withdraw power from the power-grid system to ensure any fast deviation from the nominal frequency is quelled [19]. These systems act in a manner of seconds in, for example, Continental Europe. They typically react in a manner of seconds and should stop the growth of frequency deviations in under a minute. This control mechanism is denoted, in the engineering jargon, primary control. In a language closer to physics, this mechanism ensures only that this large dynamical system finds a new stable fixed point of operation—yet not necessarily at the desired nominal frequency operation [20, 21]. One should not forget that coupled rotating masses can, in principle, rotate synchronously at any desired frequency. Thus, a secondary (and even tertiary) control mechanism is in place.

Commonly denoted secondary control, this is a set of longer timescale control mechanisms that are present to ensure that after a large deviation of the power-grid frequency, the system can revert back to a synchronous rotation of all coupled oscillators as close as possible to the desired nominal frequency [22]. The control actions, more precisely, the changes in the power generation of the respective power plants, are proportional to the integral of the frequency deviation. Unlike the primary control, whose job is solely to ensure any deviation is bounded, secondary control is actively the desire to take the newly obtained stable fixed point of the power-grid frequency after a perturbation and drag it back until it matches a stable fixed point at the nominal frequency. In the language of control engineering, primary and secondary control combined form a PI-controller [12].

One should note here that there are several other dimensions to the problem. Not only frequency needs to be controlled, but power flow between elements of the power-grid system as well [23]. This, although not discussed, is part both of secondary and

tertiary control. Moreover, as expected, the interaction of coupled oscillators leads also to a set of other internal oscillations in these system, denoted intra-area and inter-area oscillations, referring to their local and more global aspects in a power grid, that need to be managed [24].

Lastly, what is here discussed revolves solely around the operation of a power-grid system, i.e., a set of coupled oscillators, around an already stable fixed point [25]. One should mark that more catastrophic events are possible in power-grid systems, namely partial or total blackouts [26]. These are complete losses of stability of the system, which, picking up on the language on network science, are equivalent to losses of connectivity in these networks [27]. These, mind, can be both physical, in a sense that a transmission line is broken, as well as simply the decoupling of generators, thus not necessarily stemming from a physical change. These, thankfully, happen rarely [28].

Modern power-grid systems begun, from the beginning of this millennium, undergoing a fundamental transition from conventional, fossil based generation, to renewable sources of energy to mitigate climate change [29, 30, 31, 32, 33]. Maybe the most pronounced one is the change of centralised to de-centralised power generation [14]. Traditional power-grid systems are roughly based on a concept of centralised production of energy, at a large power plant, from which power is distributed, first over a long-distance transmission lines, next to regional, and lastly to local distributions grids. Due to the technological advances on both wind turbine and photo-voltaic technology, as of this decade, it is now possible to generate substantial amounts of power from a single generator of this type. This however yields an impressive control problem: the scattering of thousands to millions photo-voltaic and wind turbine units in a grid, whereas the number of dispatchable generators decreases.

The change in power generation sources implies that the traditionally vertically designed power-grid systems, where power flows from a large producer to a swat of numerous local consumers, now sees power flowing from the bottom up. Every local producer has the ability—and desire—to sell their energy. This, particularly for the study in this thesis, adds an element of uncertainty not present in a similar scale before: irregular and unpredictable fluctuations in power generation. The characteristics of these fluctuations are augmented by the nature of the energy sources, i.e., their volatile power generation [34, 35]. Unlike conventional power plants, as nuclear or coal-based power plants, which can deliver a steady and controllable amount of power, renewable energies are plagued with uncertainty. Wind turbine power generation is entirely dependent on weather conditions: a lack of wind flow implied a lack of power generation [36, 37].

Likewise, solar power generation, however certain about the time of the day the sun is shining, is subject to the movement of clouds [38]. This makes solar power generation as uncertain and volatile as wind power generation.

Compounding this problem is the fact that most renewable energy sources do not share the same intrinsic relation discussed above between power generation and angular velocity of a synchronous rotating machines [39]. As mentioned above, the inertial rotation of synchronously rotating masses plays a crucial role in ensuring power-grid frequency is kept at a desired nominal angular velocity [40]. Wind and solar power generation do not possess any intrinsic inertia. Particularly, solar photo-voltaic is a power generation procedure without any “moving parts” and thus has no rotational inertia. Most photovoltaic power sources are connected to the grid via simple power-electronic inverters, which simply follow the grid’s voltage and frequency. Advanced inverter concepts are being developed that strives to mimic the physical relation of power generation and frequency of synchronous machines to contribute to the stability of the grid. [41, 42].

On a broader spectrum it is conjectured that the ongoing increase of renewable energy sources of energy felt across the globe will lead to increased fluctuations in power-grid systems. This is certain at the level of power generation, as it proves to be a problem already existent. From the point-of-view of power-grid frequency, this rationale is not straightforward [43]. A reduced amount of inertia in a power-grid system, i.e., a smaller amount of rotating masses overall, leads to the immediate realisation that stricter control is needed, to ensure large excursions of the power-grid frequency are quelled within the agreed allotted time [44, 40]. On the other hand, strict control measures are already in place, which account for this change of status quo. Nevertheless, this is a pressing issue for control actions in power-grid systems, as a stable power-grid system is paramount to the functioning of modern society.

Augmenting this is the presence of fluctuations in power-grid frequency, which, to this date, has not seen thorough scientific examination. High-frequency ambient oscillations, i.e., overall stochastic fluctuations, are ubiquitous in power-grid frequency recordings [45, 46]. A stochastic element, i.e., what one denoted “noise”, is present in any physically driven system. This noise element can be viewed as more than just an additional nuisance in real-world recordings. Thus, instead of filtering it out, one can examine it, for it can carry some of the more interesting properties of the underlying physical process. This is the perspective taken throughout this thesis. But in the context of power-grid frequency what can this noise element represent? This noise is, interest-

ingly, an agglomeration of phenomena [47]: It comprises local properties of the location where a recording is taken; it comprises interactions with (spatially) close fluctuations, which are possibly due to local effects that propagate to nearby areas, or specific changes in generation, consumption, faults, etc., that affect the adjacent areas. Note here that certainty about the actual origin and specific elements that generate the stochastic fluctuations in power-grid frequency is not granted. However, this is not concerning. What can concern scientific investigation is what one can learn from the characteristics of these stochastic fluctuations. In particular, what one can learn from power-grid frequency in relation to other physical phenomena, i.e., what are the statistics of the stochastic fluctuations? Are they correlated to some discernible underlying phenomenon? How do they grow with external factors? And in comparison between different recordings?

Yet, there are events that one can clearly pin-point in power-grid frequency with events on energy systems. Notably, the aforementioned market activities, which involve a set of producers selling power on fixed time slots, in specific at each 15 minutes in Continental Europe, leads to large deterministic deviations in the power-grid frequency. Actually, the presence of large control mechanisms serves both as a counter mechanism to these known and scheduled market activities, as well as sporadic unexpected changes (e.g. power line failures). These deviations lead to large excursions from the nominal frequency, which can be distinctly seen each 15 minutes [48].

Moreover, one can distinctly see what is commonly described in the mathematical sciences as “mean reversion”. This is not surprising. What one observes in power-grid frequency recordings are small fluctuations around the nominal frequency, i.e., small excursions away from this nominal frequency, which revert back to the nominal value due to the elaborate control system and synchrony across the power grid. Moreover, large excursions do occur—large here can be understood as being considerably larger than the expected variance of a mean-reverting stochastic process, i.e., in this thesis, an Ornstein–Uhlenbeck process [49]. Even when these large deviations do occur, this is a deterministic phenomena, as if one drives the frequency away from the nominal value. This, as discussed, is equivalent to temporarily obtaining a new fixed stable point, or in the language on stochastic processes, the mean-reverting drift term is temporarily changed. Even in these cases one observes the strict phenomenon of mean reversion, where the frequency fluctuates around a moving drift value.

Examining real-world power-grid timeseries is key to understanding the characteristics, functioning, and potential problems in power-grid systems. Foremost, as is the concern of a large part of this thesis, parameter estimation is possible under a stochastic

description of power-grid frequency. Secondly, a description of power-grid frequency as a stochastic process can be derived from first-principals in a similar fashion to a dynamical systems’ approach, yet no clear description—or even explanation—is possible to offer to the nature of the stochastic fluctuations. This, however, one can motivate directly from the data, utilising the various timeseries estimators applicable from stochastic process theory. Thus, one can uncover proxy parameters for the aforementioned primary control, secondary control, and the stochastic noise. Equivalently, proxy terms can uncover the market activities behind the generation and consumption of energy. Lastly, embedded in the physical nature of the “fluctuations of the fluctuations”, rich phenomena of the strength of diffusion, the propagation of fluctuations in space, and the coupling of spatial and temporal dispersion can be uncovered. These aspects of power-grid frequency dynamics are the central phenomena discussed in this thesis, as evidenced in the scientific publications it comprises.

1.2 Paleo-climatic transitions and Dansgaard–Oeschger events

Understanding paleo-climatic events is fundamental to understanding today’s climate as well as the stages of evolution of the Earth’s atmosphere, ocean, and their interaction. This naturally can only be achieved via proxy measurements, from rocks and sediments, ice sheets, corals and fossils, from which one can piece together the conditions and events of the past [50]. Particularly important for this are the recordings stored in regions of the planet not anthropically affected, as for example the heart of Greenland. Ice-core drilling in Greenland has provided the most vivid description of the recent present events, particularly for the Last Glacial Period. In these proxy recordings a surprising stamp of very fast transitions are recorded which still puzzle the scientific community, denoted Dansgaard–Oeschger events [51, 52].

Dansgaard–Oeschger events are particularly abrupt transition of the northern hemisphere temperature, seen across paleo-climatic records from the past 100 000 years [53]. These are abrupt transitions, which can seemingly result in changes of over 6° Celsius of the temperature of the northern hemisphere in a span of less than 40 years. They have, so far, only been observed in glacial periods. In particular, several proxy temperature records from the last glacial period, between circa 115 000 – 11 700 BCE, reveal roughly 25 sudden increases in the global temperature which slowly relax back to a global colder

temperature.

These abrupt transitions, affecting mainly the northern hemisphere, are particularly visible in paleo-climate records from Greenland's ice core. Particularly, the laborious efforts of the scientific ice-core drilling expeditions to the heart of Greenland led to a plethora of records. A subset of these comprises a collection of stable isotope concentration (Oxygen-18 $\delta^{18}O$, Calcium Ca^{2+} , Sodium Na^{+}) as well as dust concentration, which serve as proxy for the atmospheric and sea temperature. The concentration of oxygen-18 $\delta^{18}O$ isotope relates directly to the temperature of precipitation or evaporation of a fossil over a region, i.e., in this case Greenland. Given it is a heavier isotope than the common oxygen-16, it precipitates faster due to its heaviness, and equivalently, evaporates slower. Thus it serves as an indicator of the temperature of the water content of, or surrounding, an object or region. Of distinction are the *North Greenland Ice Core Project* NGRIP project [54, 55, 56], the *Greenland Ice Sheet Project Two* (GISP2), and the *Greenland Ice Core Project* (GRIP) [57].

These proxy records serve as the stepping stones to past events, from which one can uncover distinct phases of the Quaternary glaciation, mostly in the last 400 000 years, and in particular the ongoing ice age, where mainly in the Holocene modern human civilisation flourished on the globe. The Quaternary glaciation is permeated with colder, full-glacial (denoted stadial) and milder (interstadial) periods. The interstadial period can last several decades to millennia. Dansgaard–Oeschger events are fast transitions (in a climatic scale) between stadial and interstadial conditions.

Noticeably, there is no scientific agreement about the cause or origin of Dansgaard–Oeschger events. At the beginning of this century, some authors suggested these events are periodic, roughly happening every 1 470 years [58], yet a view that these events are Poisson distributed seems more likely [59, 60]. More recent hypothesis, also backed by an agreement that Dansgaard–Oeschger events are world-wide effects, put weight on a coupled effect between ocean-atmosphere interaction. In particular, Dansgaard–Oeschger events could be coupled with changes of the Atlantic meridional overturning circulation [61], which manifests in a change in the mixing of the southern and northern water currents. The exact causal relation between Dansgaard–Oeschger events and changes in the Atlantic meridional overturning circulation is still unknown.

One interesting task is the examination of paleo-climate records under the purview of data-driven stochastic models [62, 63]. In particular, due to the clear presence of two distinct states in the paleo-climate proxy records of the global temperature, stochastic models with bistable potentials have been proposed [64]. These models often are of-

ten pre-designed, i.e., carry already a set of underlying assumptions (e.g. Markovianity, memory components, delayed coupling), but have so far not included explicit discontinuous trajectories. Moreover, the explicit functional forms of the drift or mean reverting terms are given *a priori*, and a subsequent best-fitting parameter extraction is sought. This thesis discusses a particular non-parametric estimation of the parameters underlying the stochastic process driving the oxygen-18 and dust records, under the particular case of jump-diffusion processes.

1.3 Stochastic processes

This thesis is centred on continuous-time Markovian stochastic processes, yet not necessarily continuous processes. An extension of commonly employed Langevin-like stochastic processes with a jump component is presented as

$$dX(t) = a(x)dt + b(x)dB(t) + \xi dJ(t), \quad (1.1)$$

where $a(x)$ is the drift strength, $b(x)$ is the diffusion or volatility, $B(t)$ is a Brownian motion, and $J(t)$ is a time-homogeneous Poisson jump process with rate $\lambda(x, t)$ and an amplitude ξ that is normally distributed as $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ [49]. This thesis is centred on this particular choice of Poissonian jumps with Gaussian distributed weights, but the aforementioned model is extendable to other discontinuous types of jumps, e.g. Lévy or Beta-distributed processes [65]—naturally caring for the proper derivation of the relation between moments and parameters of the processes [66]. This model, in this specific formulation, was initially proposed in Ref. [47], derived in Refs. [67, 38], and extended to provide a formal derivation of the impact of discretisation as well as a criterion to discern between pure diffusions and jump-diffusion processes in Ref. [68].

1.3.1 Applications in power-grid frequency studies

For the study of power-grid frequency dynamics, an elementary Langevin-like approach is offered [49]. More specifically, an Ornstein–Uhlenbeck process is employed as a surrogate model for power-grid frequency is proposed, i.e., $X(t) = \omega(t) = 2\pi f(t)$, where $f(t)$ is the commonly used frequency minus the reference frequency of the grid. The drift coefficient $a(X)$ always features a mean-reversion term $a = -\theta X(t)$ and possibly other deterministic contributions. Furthermore, $b(x) = \sigma$ is the volatility or diffusion coefficient, and $B(t)$ is a Brownian motion or Wiener process. The jump-like elements are not present in power-grid models.

The model was motivated in Ref. [69], where it was proposed for the dynamical system’s model for a aggregated “bulk” angular frequency in Ref. [44] by incorporating an *ad hoc* stochastic uncorrelated “noise” element. In particular, the model is an extension of a simple swing equation, i.e., a second-order ordinary differential equation, augmented with a stochastic noise. This approach equates the mean-reverting term θ in $a(x)$ to the aforementioned primary control, a second term proportional to in the integral over the frequency to be the secondary control, and a deterministic driver serving as a proxy for the imbalance of power. The stochastic noise of the process is added *ad hoc* with the desired complexity, i.e., so far solely uncorrelated Gaussian noise.

1.3.2 Applications in paleo-climate studies

In the context of paleo-climate studies the application of a bivariate jump-diffusion process is put forth. In particular, a data-driven derivation under the purview of discontinuous stochastic processes is offered for the oxygen-18 isotope $\delta^{18}O$ and the dust recording for the last glacial period, roughly 120 000 to 10 000 years Before the Common Era (BCE), relating to Dansgaard–Oeschger events in this period. In this format, no explicit functional form for the drift $a(x)$ or the diffusion $b(x)$ is chosen *a priori*, these are extracted via non-parametric estimators. The presence of discontinuous transitions, i.e., the jump components, is motivated via a study of higher-order Kramers–Moyal coefficients and their scaling, and previous observation suggesting Dansgaard–Oeschger events are Poisson distributed.

1.4 Overview of relevant publications

This thesis comprises eight scientific publications, five of which are published, one submitted, and two in preparation. These are:

- #1 **[published]** L. Rydin Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, and B. Schäfer. *Data-Driven Model of the Power-Grid Frequency Dynamics*. IEEE Access **8**, 2020, pp. 43082–43097, Ref. [1].
- #2 **[published]** M. Anvari, L. Rydin Gorjão, M. Timme, D. Witthaut, B. Schäfer, and H. Kantz. *Stochastic properties of the frequency dynamics in real and synthetic power grids*. Physical Review Research **2**(1), 2020, p. 013339. Ref. [2].

- #3 **[published]** L. Rydin Gorjão and F. Meirinhos. `kramersmoyal`: Kramers–Moyal coefficients for stochastic processes. *Journal of Open Source Software* **4**(44), 2019, p. 1693, Ref. [3].
- #4 **[published]** L. Rydin Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer. *Open data base analysis of scaling and spatio-temporal properties of power grid frequencies*. *Nature Communications* **11**, p. 6362, 2020, Ref. [4].
- #5 **[in preparation]** L. Rydin Gorjão, L. Vanfretti, D. Witthaut, C. Beck and B. Schäfer, under the working title *Phase and amplitude synchronisation in power-grid frequency fluctuations*, Ref. [5].
- #6 **[published]** L. Rydin Gorjão, J. Heysel, K. Lehnertz, and M. R. R. Tabar. *Analysis and data-driven reconstruction of bivariate jump-diffusion processes*. *Physical Review E* **100**, 2019, p. 062127, Ref. [6].
- #7 **[submitted]** L. Rydin Gorjão, D. Witthaut, and P. G. Lind. *JumpDiff: A Python library for statistical inference of jump-diffusion processes in sets of measurements*, submitted to the *Journal of Statistical Software*, 2020, Ref. [7].
- #8 **[in preparation]** L. Rydin Gorjão, K. Riechers, F. Hassanibesheli, D. Witthaut, and P. G. Lind, under the working title *Dansgaard–Oeschger events: Change in stability and jumps modelled via univariate and bivariate jump-diffusion processes*, Ref. [8].

Part of this Doctoral thesis was also the making of Refs. [70, 71, 72] and the software in Ref. [73], but these four scientific works are not included in this thesis.

Publication #1: Data-driven model of the power-grid frequency dynamics

In this publication an easy-to-use, data-driven stochastic model is designed to generate exemplary trajectories of power-grid frequency. In the same manner, a set of non-parametric estimators is offered to retrieve from data the primary and secondary control mechanisms ensuring balance in power-grid frequency dynamics. A second-order stochastic differential equation is presented, where the short-term primary control (acting in a scale of a few seconds) is retrieved by extracting the drift coefficient $a(x)$. The long-term recovery control, the secondary control, is recovered by the deterministic exponential decay of the power-grid frequency trajectories. Intrinsic stochastic noise is retrieved through

the diffusion terms $b(x)$. Lastly, the flexibility of the model lies in the ability of adding a deterministic driving function for the power-grid frequency. The balance of power generation and consumption is, from the point-of-view of a system operator, a deterministic driver function, where the difference should vanish. Excess or lack of power results in large deterministic deviations of the power-grid frequency.

The publication offers an explanation of why the statistics of power-grid recordings in Continental European are always platykurtic, and conversely the Great British power-grid frequency recordings are leptokurtic, just as it is possible to do so with any deterministically driven stochastic system.

Publication #2: Stochastic properties of the frequency dynamics in real and synthetic power grids In this publication an extended examination of power-grid frequency recordings is offered for Continental Europe and Great Britain, for the years 2015, 2016, and 2017.

Taking an Markovian Ornstein–Uhlenbeck model as the basis of the analysis, a characteristic damping constant is obtained by analysing the autocorrelation functions of the data, yielding the relaxation time of power-grid frequency in both power grids. A study of the increments of power-grid frequency is put forth, where a super-statistical q -Gaussian (with $q \neq 1$) is shown to fit the distribution of the increments. A study of the power spectrum is presented to justify the stationarity of the recordings, two three-point correlation functions are offered to justify time-symmetry in the data, and an examination of the Chapman–Kolmogorov test yields that the processes are Markovian. This early analysis of power-grid frequency records features a set of simplifying assumptions, which are critically reviewed in Section 3.1.

Publication #3: kramersmoyal: Kramers–Moyal coefficients for stochastic processes In this publication an N -dimensional non-parametric Nadaraya–Watson estimator of the Kramers–Moyal coefficients and conditional moments of stochastic time-series, to any desired order, is presented. The software is based on numerically efficient convolutional procedures in the computer language Python. The software was used extensive in publications #1, #4, #6, and #7.

Publication #4: Open data base analysis of scaling and spatio-temporal properties of power grid frequencies In this publication a study of the statistical properties of an extensive data collection of power-grid frequency recordings is put forth.

The statistical properties of the distribution of the frequency in various grids around the globe is presented and a scaling of the diffusion coefficient $b(x)$ first presented in Ref. [69] is confirmed. A subsequent analysis of spatio-temporally distributed recordings in Continental Europe in 2019, comprised of six synchronised measurements, reveals the existence of strong correlations of the increments of power-grid frequency time series. A study of a relaxation time for synchronisation is also presented, yielding a diffusion-like relation between relaxation time and squared-distance, resulting in a first examination of the relaxation of fluctuations in power-grid frequency dynamics, which seems to follow a diffusive behaviour.

Publication #5: Phase and amplitude synchronisation in power-grid frequency fluctuations In this publication a thorough examination of six high-frequency synchronous power-grid frequency recordings from the Nordic synchronous area is put forth. A separation of phase and amplitude synchronisation is proposed based on the distinction of temporal correlation and fluctuation relaxation of the incremental time-series of the power-grid recordings. Thus the scale of phase synchronisation is found to take place < 2 second scale whereas the amplitude synchronisation takes place in $2 \sim 5$ seconds. Moreover, phase synchronisation emerges in a linear relation in the spatial separation of recordings whereas amplitude synchronisation emerges diffusively, as first observed in Publication #4. Additionally, it is posited that the class of diffusive amplitude synchronisation falls in the category of a super-diffusive process, intrinsically linked with the temporal correlations of each timeseries, i.e., to their Hurst coefficient.

Publication #6: Analysis and data-driven reconstruction of bivariate jump-diffusion processes In this publication bivariate jump-diffusion process are introduced, alongside a data-driven, non-parametric estimation procedure of higher-order (up to 8) Kramers–Moyal coefficients, allowing the reconstruction of all relevant parameters in a jump-diffusion process. The procedure is validated numerically, presenting the limitations in the presence of coupled processes, the capability of retrieving the jump elements, and the numerical limitations for short timeseries.

Publication #7: JumpDiff: A Python library for statistical inference of jump-diffusion processes in sets of measurements In this publication a Python library denoted JumpDiff is presented, comprising of non-parametric estimators to retrieve a relevant parameters of one-dimensional jump-diffusion processes. The software relies

of the mathematical methods derived in Publication #3, specialised for jump-diffusion processes in one dimension. Furthermore presented is a set of second-order corrections to the Kramers–Moyal operator, represented as the solution of the Kramers–Moyal equation for discontinuous Markovian stochastic processes via an exponential representation and approximation of the Kramers–Moyal operator, extending the work in Ref. [74]. The software includes as well a criterion to discern between pure diffusions and jump-diffusion processes, following the basic methods introduced in Ref. [68].

Publication #8: Dansgaard–Oeschger events: Change in stability and jumps modelled via univariate and bivariate jump-diffusion processes In this publication an analysis of Dansgaard–Oeschger (D–O) during the Last Glacial Period under the purview of bivariate jump-diffusion processes is presented. A data-driven analysis of $\delta^{18}O$ and dust recordings suggests that there is a change from a bistable to a unstable potential of the dust count, via an imperfect supercritical pitchfork bifurcation.

Furthermore, the $\delta^{18}O$ recording is discontinuous and thus best modelled via a jump-diffusion model. The aforementioned criteria to discern between continuous and discontinuous stochastic processes is employed to separate the stochastic nature of the $\delta^{18}O$ and dust recordings. Lastly, the coupling of any terms in the bivariate jump-diffusion is shown to be vanishing, suggesting that the D–O are deterministically triggered.

Chapter 2

Publications

2.1 Stochastic analysis and modelling of power-grid frequency dynamics

2.1.1 Publication #1

L. Rydin Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, and B. Schäfer. *Data-Driven Model of the Power-Grid Frequency Dynamics*. IEEE Access **8**, 2020, pp. 43082–43097, Ref. [1].

Status: published

Received November 22, 2019, accepted December 19, 2019, date of publication January 20, 2020, date of current version March 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967834

Data-Driven Model of the Power-Grid Frequency Dynamics

LEONARDO RYDIN GORJÃO^{1,2}, MEHRNAZ ANVARI³, HOLGER KANTZ³,
CHRISTIAN BECK⁴, DIRK WITTHAUT^{1,2}, MARC TIMME^{5,6},
AND BENJAMIN SCHÄFER^{4,5,6}

¹Forschungszentrum Jülich, Institute for Energy and Climate Research—Systems Analysis and Technology Evaluation (IEK-STE), 52428 Jülich, Germany

²Institute for Theoretical Physics, University of Cologne, 50937 Cologne, Germany

³Max-Planck Institute for the Physics of Complex Systems (MPIPKS), 01187 Dresden, Germany

⁴School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, U.K.

⁵Chair for Network Dynamics, Center for Advancing Electronics Dresden (cfaed), Institute for Theoretical Physics, Technical University of Dresden, 01062 Dresden, Germany

⁶Network Dynamics, Max Planck Institute for Dynamics and Self-Organization (MPIDS), 37077 Göttingen, Germany

Corresponding author: Leonardo Rydin Gorjão (l.rydin.gorjao@fz-juelich.de)

This work was supported in part by the Federal Ministry of Education and Research (BMBF) under Grant 03SF0472 and Grant 03EK3055, in part by the Helmholtz Association (via the joint initiative *Energy System 2050 - A Contribution of the Research Field Energy*, under Grant VH-NG-1025, in part by the German Science Foundation (DFG) by a grant toward the *Cluster of Excellence Center for Advancing Electronics Dresden* (cfaed), in part by the *STORM - Stochastics for Time-Space Risk Models* project of the Research Council of Norway (RCN) under Grant 274410, and in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie under Grant 840825.

ABSTRACT The energy system is rapidly changing to accommodate the increasing number of renewable generators and the general transition towards a more sustainable future. Simultaneously, business models and market designs evolve, affecting power-grid operation and power-grid frequency. Problems raised by this ongoing transition are increasingly addressed by transdisciplinary research approaches, ranging from purely mathematical modelling to applied case studies. These approaches require a stochastic description of consumer behaviour, fluctuations by renewables, market rules, and how they influence the stability of the power-grid frequency. Here, we introduce an easy-to-use, data-driven, stochastic model for the power-grid frequency and demonstrate how it reproduces key characteristics of the observed statistics of the Continental European and British power grids. Using data analysis tools and a Fokker–Planck approach, we estimate parameters of our deterministic and stochastic model. We offer executable code and guidelines on how to use the model on any power grid for various mathematical or engineering applications.

INDEX TERMS Stochastic modelling, power-grid frequency, swing equation, control systems, parameter estimation, Fokker–Planck equation, data-driven model.

I. INTRODUCTION

The energy system is currently undergoing a rapid transition towards a more sustainable future. Greenhouse gas emissions are reduced by implementing distributed renewable-energy sources at ever growing rates in the world [1]. Simultaneously, new policies, technologies, and market structures are being implemented in various regions in the energy systems [2]. These new market structures are not necessarily benefiting the stability of the power grid: A control power shortage in the German grid in June 2019 was potentially

caused by unknown traders exploiting the energy market structure [3], [4].

The field of energy research itself is quickly developing and attracting researchers from various disciplines working towards new control systems, new market models, and new technologies every year [5], [6]. Regardless of the specific aspect of the energy system, one element remains unchanged: The electrical power system and the stability of its frequency are critical for a stable operation of our society [7].

The power-grid (mains) frequency dynamics mirrors the balance of supply and demand of the power grid: An excess of generation leads to an increased frequency and a shortage of generation leads to a reduced frequency value. The power

The associate editor coordinating the review of this manuscript and approving it for publication was Roberto Sacile.

grid is stabilised by controlling the frequency and maintaining it at a nominal frequency [8]. But the task of maintaining a set frequency across an entire power-grid system is not a simple one: systems vary in size and structure, energy sources are possibly volatile in their output, as for example are wind or photo-voltaic generators [9], [10], and the dispatch of electrical energy and market activity have an impact on the overall dynamics.

Understanding the intricacies of the frequency dynamics becomes of great importance, both to control the current power grid [8], [11] but also for implementing real-time pricing schemes [12], [13] or smart grids in the future [14]. Solid estimates of fluctuations are essential for example when dimensioning back-up or control options, such as determining the capacity of batteries or other energy storage to balance periods with highly fluctuating demand or times without renewable generation [15]. Similarly, when establishing new power grid types, such as smart grids with potentially novel electricity market structure, the market design should ideally support the stability of the grid.

While both the power-grid frequency dynamics and the stochastic nature of the power-grid frequency have been intensely studied, we require a better understanding of the interaction of frequency dynamics with both stochastic fluctuations and market behaviour. The dynamics of the power-grid variables, including frequency, voltage, reactive power, etc., may be modelled with arbitrary complexity based on various models [8], [11], [16]–[19]. Simultaneously, stochastic modelling of fluctuations within the power grid [18], [20] still often uses Gaussian noise models [14], [21], [22], while non-Gaussian statistics [9], [23] as well as deterministic events caused by trading [24] are rarely included.

Existing literature that explicitly deals with realistic forecasts of the power grid frequency often focuses on inverter control [25] or the power interface between grid layers [26]. Alternatively, forecasts are done for electricity consumption [27] or for renewable generation, such as solar generators [28]. In contrast, models that predict or even give stochastic characteristics of the power-grid frequency are very rare [29].

Here, we propose an accessible and easy-to-use stochastic model that seeks to describe the dynamics of the power-grid frequency in a reduced framework combining stochastic and deterministic factors acting on the power-grid frequency. We focus on the intermediate time scale of several seconds to few hours, leaving very short or very long time scale for future work. Simultaneously, our modelling approach balances the benefits of realistic case studies, generally applicable and abstract stochastic models as well as application-oriented data-driven approaches.

We first review the factors influencing the power-grid frequency dynamics, based on frequency recordings from European grids. Next, we introduce a general stochastic model and discuss three particular cases of how the model may be implemented. For each case we estimate the system parameters,

such as control strength and noise amplitude using stochastic theory and data-driven approaches.

We compare the frequency statistics of the models with real-world measurements to showcase how they reproduce characteristic features. Overall, our modelling approach is very flexible and easily applicable to many different power grids and could be used for planning purposes, e.g. when setting security operational limits or designing markets. We provide executable code for the model in the supplementary material.

II. FACTORS IMPACTING THE POWER-GRID FREQUENCY

To construct a model describing the intermediate time scale dynamics and characteristics of the power-grid frequency, we must first recall the nature and the intricate details of the power-grid frequency dynamics, both deterministic and stochastic, as we observe them in frequency trajectories [30], see Fig. 1.

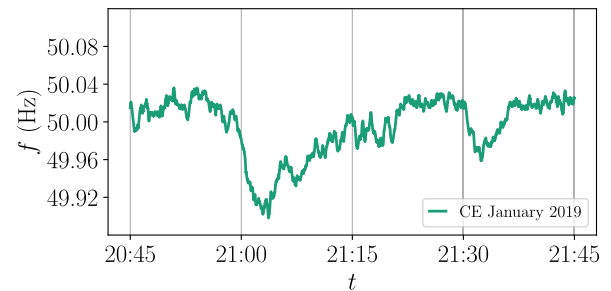


FIGURE 1. The frequency dynamics is influenced by both stochastic and deterministic aspects. The trajectory of the power-grid frequency is substantially influenced by stochastic effects, as seen by the erratic motion. In addition, we observe deterministic behaviour: Every 15 minutes (vertical lines) the frequency abruptly decreases and then slowly trends upwards for the next 15 minutes. The plot uses the TransNetBW data [30] from the European Central power grid CE, from January, 10th 2019, 20:45 to 21:45.

The power-grid frequency is not following a simple Gaussian process but displays heavy tails and regular correlation peaks, see Fig. 2 and [23], [31], [32] for more detailed analysis. To get a better understanding of the different factors impacting the grid frequency, we give an overview of these: First, we review the innate and humanly devised control systems, continue with the market and power dispatch design and close the section with a stochastic description of the noise acting on the power grid.

A. THE FUNDAMENTAL CONTROL SCHEMES

The power supply of the grid is designed so that the frequency of the alternating current is kept steadily at a fixed nominal value, i.e., 50 Hz in Europe and many parts of the world, or 60 Hz in the Americas, Southern Japan and some other regions. The electrical frequency of e.g. 50 Hz corresponds to large mechanical generators rotating in synchrony at this frequency (or integer multiples of it) across each synchronous region, such as the Continental European grid. How is this

frequency kept fixed when facing fluctuations or larger disturbances?

Suppose a large generator disconnects from the grid while the power demand in the region stays constant. The missing energy cannot be drawn from the grid itself, as it cannot store any energy directly [34]. Instead, power is first provided by inertial energy until primary, secondary, and potentially tertiary control set in to ensure the provision of the missing power [34]. In the first moments after the disturbance, the missing power is drawn from the kinetic energy of the large rotating machines. Their kinetic energy is converted into electrical energy and the generators are slowed down, thereby reducing the overall frequency in the grid. This *inertial response* ensures the system does not drift off from its designed nominal frequency too rapidly and smoothen any disturbances. Nevertheless, the generators continue to slow down. Moments later, *primary control* activates: Dedicated power plants, and recently also battery stacks [35], measure the deviation of the frequency from the reference and insert additional power into the grid proportional to the frequency deviation. This power influx prevents a further decrease of the frequency and stabilises it at a fixed but lower frequency, which is not desired for operation, as any further problems might cause the frequency to leave the stable operational limits [8], [34]. While the primary control compensates for the missing power, the kinetic energy of the rotors is still lower than initially and thereby the frequency is not at the reference value. To restore the frequency back to the reference frequency an integrative control, *secondary control*, is necessary. A few minutes after the disturbance, this control fully restores the energetic state and the grid is brought back into a new stable state at its nominal frequency (i.e., 50 Hz or 60 Hz, depending on the grid in question). On even longer time scales of potentially hours, *tertiary control*, often operated manually, sets in [36]. As this tertiary control sets in, primary and secondary control can be reduced to become available for further control actions.

Here, we focus on the effects of inertia, as well as primary (proportional) and secondary (integrative) control in our synthetic model. The time scales of these three controls are significantly different, and they functionally react to deviations of different variables of the system: Where primary control stabilises the grid based on the frequency deviations of the system, the secondary control balances the total power to ensure stability based on an integral of the frequency, i.e., an angle.

As a recent challenge, the replacement of conventional power generators with renewable generators reduces the overall system inertia [37] and thereby makes complementary control mechanisms or virtual inertia increasingly important [38].

B. ELECTRICITY DISPATCH AND MARKET

While the control schemes keep the frequency close to the reference for small and unforeseen changes of supply and demand, an electricity market has been established to

coordinate longer-term power dispatches dealing with large and predictable variations.

The effective demand acting on the power grid is the aggregation of millions of consumers throughout the synchronous region. This aggregated demand is continuously changing over time since consumption during the day tends to be higher than during the night and industrial activities during the week lead to higher consumption than during the weekends [34].

The continuously changing demand has to be met with sufficient supply of electrical power in the same synchronous grid. Therefore, power plant operators have to adjust their generation according to the needs of the consumers. While some power plants, such as gas turbines, can ramp their generation up or down very fast, other plants, such as coal or nuclear power plants, require more time and therefore prefer to commit generation for longer time periods [36], [39]. Demand response schemes, where consumers shift their demand to periods of higher generation, bring additional flexibility to the grid [40].

To reach an economic optimum on who is supplying and when, power-plant operators bid on spot markets to offer power generation [36]. This includes a *day-ahead* market to fulfil the expected power demand, and an *intra-day* market acting on time scales of few hours to several minutes, to balance short-term mismatches, amongst other [41]. This bidding on the market takes place in discrete time-slots: Any power provided by one operator is provided for a fixed interval, e.g. one hour, half an hour, or 15 minutes, as is often the case, such as in the European Energy Exchange (EEX) [42].

An important consequence of the fixed intervals of generation is that it does not perfectly fit the smooth demand curve. If we approximate a monotonically increasing demand function (such as during the early morning hours) with a step function assuming the mean for a given time interval, we will initially overestimate the demand, which is still growing. After some time, supply and demand perfectly match but then the demand surpasses the supply again. This leads to the balance between supply and demand being approximately a sawtooth function, see Fig. 3.

Indeed, we also observe the consequences of the intervals when analysing the frequency trajectory [24] or its autocorrelation in the Continental European grid. The frequency displays regular surges and sags approximately every 15 minutes, where the supply updates to the new demand interval. At full hours these effects are more pronounced since the total dispatch and trading volume is higher at full hours compared to other 15 minute intervals [43]. Not only the frequency trajectory displays these jumps and sags, see Fig. 1 but also the autocorrelation function of the power-grid frequency $c(\Delta t)$ reveals distinct peaks at 15, 30, 45 and 60 minutes, see Fig. 2 and [23], [31].

We will include the market influence by employing a deterministic power-mismatch model in our stochastic model. But more importantly, we can extract vital information by observing this phenomenon, as we will highlight below.

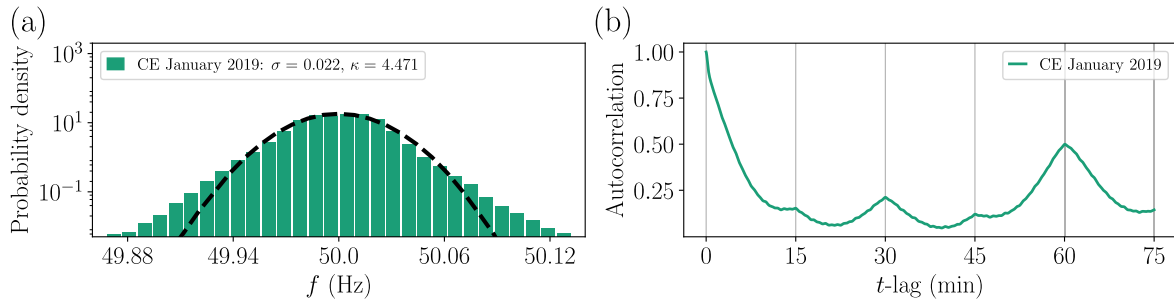


FIGURE 2. The power-grid frequency is heavy-tailed and has regular correlation peaks. (a) The frequency histogram displays heavy tails, which are quantified by a kurtosis κ that is much larger than the Gaussian value of $\kappa_{\text{Gaussian}} = 3$. Consistently, the best-fitting Gaussian distribution (dashed line) does not capture the tails. (b) The autocorrelation function of the grid frequency decays exponentially within the first minutes, which is a typical behaviour for many stochastic processes [33]. In addition, the autocorrelation peaks every 15 minutes due to trading activity. The plots use the TransNetBW data from January 2019 [30].

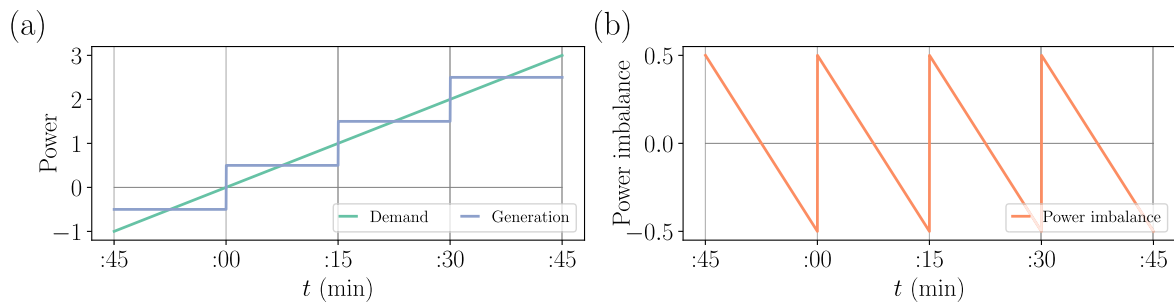


FIGURE 3. The effective power balance approximates a sawtooth function over time. We schematically depict the interplay between generation, demand and the resulting power imbalance: (a) While the demand curve is approximately smooth, the scheduled generation approximates the curve using step functions. (b) The resulting power balance is approximately a sawtooth function with jumps upwards and ramps downwards if the demand rises and ramps upwards and jumps downwards if the demand decreases. Here, we display all jumps with the same height for simplicity. In our model, we use different jump heights of the Heaviside and thereby also of the sawtooth function for hourly, half- or quarter-hourly jumps.

C. NOISE

So far, we have introduced the two deterministic elements of our model: Control in the form of inertia, primary and secondary control, and electricity trading occurring at fixed times. We are only missing the stochastic element of the model, i.e., the *noise* acting on the system. Noise here is meant as any form of stochastic fluctuation. Its sources are plentiful, ranging from demand fluctuations [40], [44] to intermittency in the renewable generators [9], [45], thermal fluctuations, and others, many of which are typically unknown [23]. However, the precise origin of the noise is not essential for our modelling approach. In fact, we only observe the cumulative effect of the noise in how it influences the power-grid frequency, regardless whether it originates from local disturbances or system-wide variations. Aggregating all sources of noise allows it to be handled as a stochastic process, see also [32] for more details.

As a first approximation for the noise, we will assume white Gaussian noise, based on two important observations. First, Gaussian noise arises naturally in many settings due to the *Central Limit Theorem*. In its simplest form it states that the sum of randomly drawn numbers, in our case the aggregation of renewable, demand and any other form of fluctuation, approximates a Gaussian distribution if sufficiently many

contributions are summed up [33]. Second, we note that non-Gaussian frequency distributions can easily be described by super-imposed Gaussian distributions, following *superstatistics* [23], [38], [46], where parameters, such as the standard deviation change over time. Moreover, the above mentioned trading intervals are known to contribute significantly to these tails [31].

If so desired, employing another form of noise is left open in the model, without any fundamental change of the model itself. There are plenty of non-Gaussian sources of noise impacting the power grid, such as jump noise from solar panels [9] or turbulence from wind turbines [20], [47]. Instead of Gaussian noise, we could include for example non-Gaussian effects via Lévy-stable distributions or q -Gaussian distributions [38], [48].

III. DATA-DRIVEN MODEL

Now, we formulate a simple dynamical model for the frequency dynamics that includes all factors influencing the power-grid frequency. First, we present the model and explain how the above-mentioned factors enter the model. We then discuss special cases of how some parameters could be set as constants or as time-dependent. We close the section by proving the theory to estimate the parameters of the model.

For simplicity, we do not use the frequency f as the variable but the bulk angular velocity $\omega = 2\pi(f - f_{\text{ref}})$, with reference frequency $f_{\text{ref}} = 50$ or 60 Hz, i.e., we move into the rotating reference frame. In this frame, the dynamics of the angular velocity ω and the bulk angle θ may be modelled in an aggregated swing equation [49] as

$$\begin{aligned} \frac{d\theta}{dt} &= \omega, \\ M \frac{d\omega}{dt} &= -c_1\omega - c_2\theta + \Delta P + \epsilon\xi. \end{aligned} \quad (1)$$

The factor M gives the inertial constant of the system and sets the time scale it reacts to changes. For simplicity, we absorb it in the remaining constants and set $M = 1$ in the following, i.e., $c_1 \rightarrow c_1/M$, $c_2 \rightarrow c_2/M$, $\Delta P \rightarrow \Delta P/M$ and $\epsilon \rightarrow \epsilon$.

The term $-c_1\omega$ models primary control and general damping acting on the system [16], [34]. The larger the deviation from the nominal frequency, i.e., the larger ω , the larger the damping and control force.

The expression $-c_2\theta$ models the secondary control [50], [51]. If the system deviates from the nominal frequency, e.g. because $\omega > 0$ for a long time, then the bulk angle θ increases more and more and thereby the secondary control increases and acts as an increasing force to return the system towards the nominal frequency. We use the simplest integral control, whereas other secondary control implementations [50], [52]–[56] might be considered in the future. Typically, the magnitude of the primary control parameter is much larger than the secondary control parameter $c_1 \gg c_2$ to implement that primary control acts faster than secondary control.

The power mismatch is given as ΔP . It contains only the deterministic mismatch between supply and demand. If generation surpasses consumption, ΔP becomes positive and vice versa. In our market model, we will employ a time-dependent ΔP , inspired by empirical power trajectories, see Fig. 3.

Finally, $\epsilon\xi$ denotes the aggregated noise acting on the system. As pointed out in the previous section, we assume ξ to be white Gaussian noise, i.e., its time average is zero $\langle \xi(t) \rangle = 0$ and its correlation is zero for non-identical times, i.e., it is a delta function $\langle \xi(t)\xi(t') \rangle = \delta(t - t')$ [33]. Extensions using correlated or non-Gaussian noise are also possible in the same framework and can prove very useful if the noise function or its characteristics are known. From an *a priori* point-of-view, employing coloured noise would required an explicit knowledge of its presence in power-grid frequency systems and would complicate the parameter extraction.

The model (1) is very general as we have not yet specified the parameters c_1 , c_2 , ϵ or the function ΔP . Note again the different roles of primary and secondary control: Assume $\Delta P = P_0 > 0$, this will increase the angular velocity ω and thereby the angle θ . Without secondary control and noise, i.e., $c_2 = \epsilon = 0$, the new quasi-steady state becomes $\omega^* \approx P_0/c_1 > 0$. The full fixed point $\omega = 0$ can only be restored with an additional (integrative) secondary control.

A. CASES

We consider some special cases of parameter choices for model (1) here. Theoretically, the model proposed so far would allow that the three parameters c_1 , c_2 , and ϵ are chosen as zero or non-zero constants, time-dependent functions, or to follow their own stochastic process. Similarly, the power mismatch ΔP could be any function, as long as the differential equation is still well-defined. We review three cases, see also Fig. 4 for an overview.

The distinguishing factor between those cases is the role of secondary control c_2 and power imbalance ΔP : Any non-zero power imbalance ΔP will be compensated by secondary control if $c_2 > 0$. This means from a data-analysis it is virtually impossible to distinguish cases where $\Delta P = 0$ and no secondary control is active or $\Delta P \neq 0$ and secondary control restored the frequency or a case where a slowly changing ΔP restored the frequency on its own without secondary control active. Complementary, large and rapid changes in the power imbalance are clearly visible in the frequency trajectory and always have to be included in the models.

Case A: A simple starting point is to set c_1 , c_2 , and ϵ all as non-zero constants. By including an active secondary control, we neglect slow changes in the power imbalance ΔP and assume that secondary control is the main restoring force following a sudden jump. Specifically, we assume that the power mismatch ΔP is given as a piece-wise constant function, i.e., a Heaviside function. This model has the advantage that we can easily estimate all parameters from the trajectory.

Case B: Alternatively, we may neglect the effects of secondary control, setting $c_2 = 0$. To balance the frequency, we then require a balanced power dispatch on average, i.e., $\langle \Delta P \rangle = 0$. A simple function to realise this, while maintaining the jumps, which are visible from the frequency trajectories, is a sawtooth function, i.e., piecewise linearly increasing or decreasing over time. Similar to Case A, we still use constant non-zero c_1 and ϵ .

Case C: We again repeat Case A but instead of estimating the power mismatch ΔP from frequency trajectories, we use historic demand data of Germany, based on data published by ENTSO-E [57].

B. ESTIMATING PARAMETERS

To generate a synthetic trajectory approximating real data, we need to estimate suitable parameters for our model. Here, we present the mathematical background and basics that allow this parameter estimation as well as illustrations of the procedure in Figs. 5, 6 and 7. We provide additional guidance and code on how the estimators can be applied in practice in the Supplemental Material.

We estimate the parameters of the synthetic model as follows: The primary control c_1 and the noise ϵ are obtained from using the first and second Kramers–Moyal coefficient respectively. Next, the power mismatch ΔP and the secondary control c_2 are determined from the trajectory at the trading times.

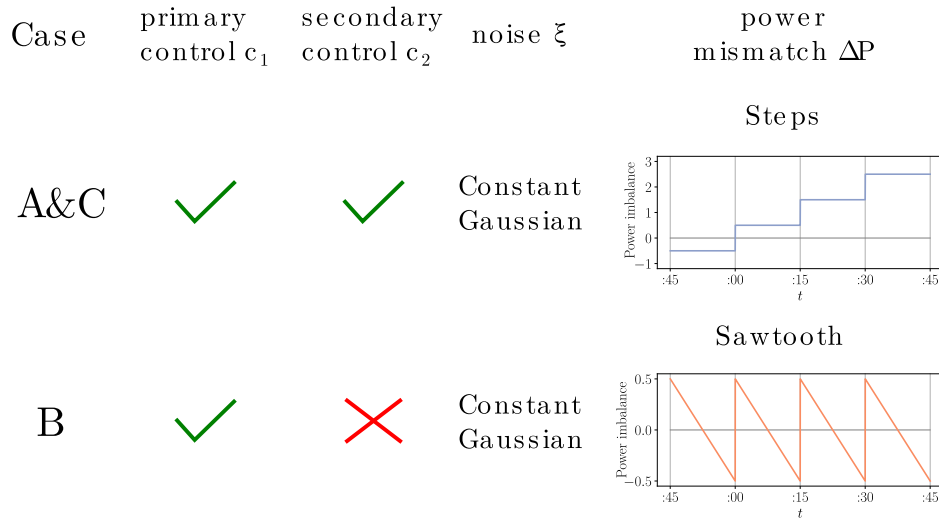


FIGURE 4. We consider three different cases to model the power-grid frequency based on the model (1). In Case A, we apply constant primary and secondary control, white Gaussian noise with constant amplitude ϵ and a Heaviside power mismatch ΔP . In contrast, Case B uses no secondary control and applies a sawtooth function for the power mismatch ΔP . We still apply a constant primary control c_1 and white Gaussian noise with constant amplitude ϵ . Finally, Case C uses Case A's settings but we extract the jump heights of the Heaviside function from independent historic demand data [57] and not from the frequency trajectory. As in Fig. 3, we display all jumps with the same height for simplicity.

1) KRAMERS–MOYAL AND FOKKER–PLANCK

Let us briefly review some relevant stochastic theory necessary to estimate the parameters. The synthetic model (1) includes stochastic and deterministic dynamics. Assuming that the deterministic contribution given by ΔP and the secondary control c_2 are either very small or subtracted from the trajectory, we are left with a purely stochastic process for ω in the form of a Langevin equation. Such an equation cannot be solved deterministically but we may formulate the Fokker–Planck equation of the stochastic system instead [33]:

$$\begin{aligned} \frac{\partial p}{\partial t} &= -\frac{\partial}{\partial \omega} (-c_1 \omega p) + \frac{\epsilon^2}{2} \frac{\partial^2 p}{\partial \omega^2} \\ &= -\frac{\partial}{\partial \omega} \mathcal{D}^{(1)} p + \frac{\partial^2}{\partial \omega^2} \mathcal{D}^{(2)} p. \end{aligned} \quad (2)$$

This Fokker–Planck equation is a partial differential equation for the probability density function $p(\omega, t)$ of the system. Solving this Fokker–Planck equation thereby returns the probability $p(\omega, t)$ to observe the system in state ω at time t , see e.g. [33] for an introduction to Fokker–Planck equations.

Terms subject to first derivatives are known as *drift terms* $\mathcal{D}^{(1)}$, while terms subject to second derivatives are called *diffusion terms* $\mathcal{D}^{(2)}$ [33]. Drift terms describe the deterministic behaviour of the full stochastic system, e.g. the movement of a particle within a potential or in our case the control and damping forces acting within the power grid, causes a “drift” towards the stable state. Complementary, the diffusion terms determine the stochastic part of the trajectory. Random noise makes state of the grid “diffuse” through the available state space and typically leads to a broadening of the probability distribution p [33]. We can read off the drift and the diffusion terms of the angular velocity ω as

$\mathcal{D}^{(1)} = -c_1 \omega$ and $\mathcal{D}^{(2)} = \frac{\epsilon^2}{2}$ respectively. These drift and diffusion terms of the Fokker–Planck equation are also known as the Kramers–Moyal coefficients from the Kramers–Moyal expansion of the fundamental master equation of the system. Only this approximation allows us to write the Fokker–Planck equation [58], [59]. From these coefficients we estimate the mentioned parameters.

From a data-driven perspective, we can recover the drift $\mathcal{D}^{(1)}$ and diffusion $\mathcal{D}^{(2)}$ coefficients strictly from the data by employing a histogram regression or a Nadaraya–Watson kernel estimator. This approach is particularly useful when the fundamental equations of motion are not known but we also use it here to approximate the functional form of the Fokker–Planck equation and recover the primary control c_1 and noise amplitude ϵ , as given by (3) and (4), respectively. The drift and diffusion coefficients are the two lowest-order Kramers–Moyal coefficients. Further information on Fokker–Planck equations, Kramers–Moyal expansion, and stochastic modelling is available in [33], [58], [59].

2) ESTIMATING THE PRIMARY CONTROL c_1

Having set out the theory of Fokker–Planck equations and Kramers–Moyal coefficients, we now apply them to determine the primary control c_1 , by applying a two-step process: We first subtract the deterministic and slow time scale components from the trajectory and then determine the first Kramers–Moyal coefficient.

We first remove the driving deterministic characteristics of the model (1) from any trajectory we analyse. To do so, we filter the data with a Gaussian kernel filtering, with a window of 60 seconds, to remove the deterministic trend and any slow process, such as secondary control, and thus

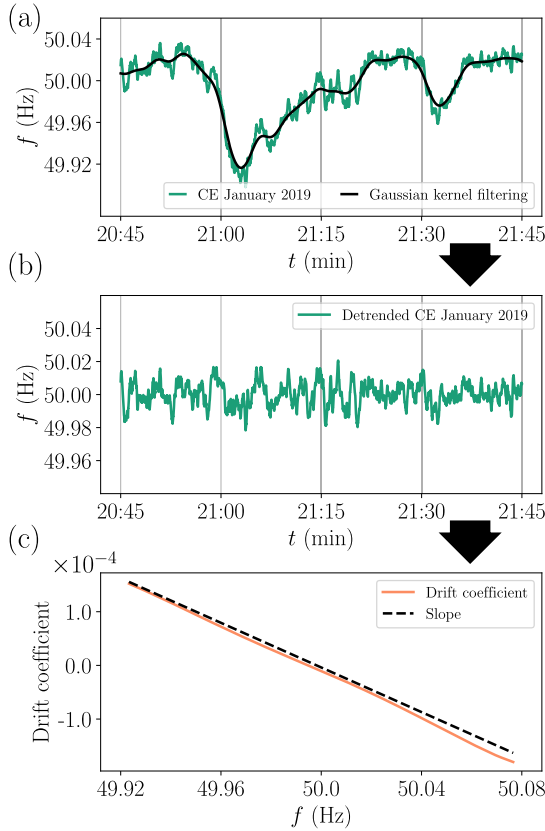


FIGURE 5. The primary control c_1 is computed from fluctuations around the trend. To estimate the primary control c_1 , we first detrend the data by applying a Gaussian kernel and then compute the drift coefficient. (a): We display a snippet of the power-grid frequency trajectory from the CE data from January 2019, as in Fig. 1, alongside with the 60 seconds window Gaussian kernel detrending, that captures the deterministic and slowly changing contributions of the power-grid frequency. (b): We extract the stochastic motion by subtracting the deterministic trend from the power-grid frequency. What is left is a stochastic trajectory resembling approximately an Ornstein–Uhlenbeck process. (c): We compute the first Kramers–Moyal coefficient, known as the drift coefficient, of the now purely stochastic process. The slope of the drift coefficient is equal to the primary control $-c_1$.

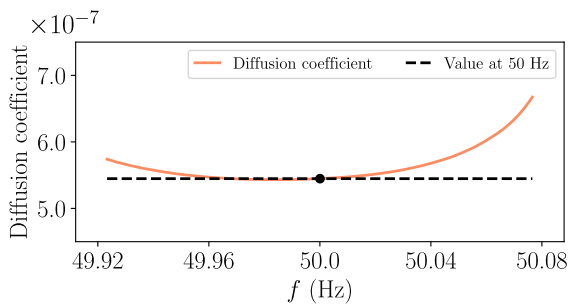


FIGURE 6. The noise amplitude ϵ is obtained using the diffusion coefficient. We display the diffusion coefficient, or second Kramers–Moyal coefficient around 50 Hz for the CE grid for the month of January 2019. By taking the value at 50 Hz, indicated on the plot, and by using relation (4), we obtain the noise ϵ .

remain solely with the stochastic component of the process. A window too small (e.g. 5 seconds) would still contain noise contributions, while a large window (e.g. 300 seconds) leads to an almost flat signal. The procedure is independent of the specific driving method (cf. Case A and Case B).

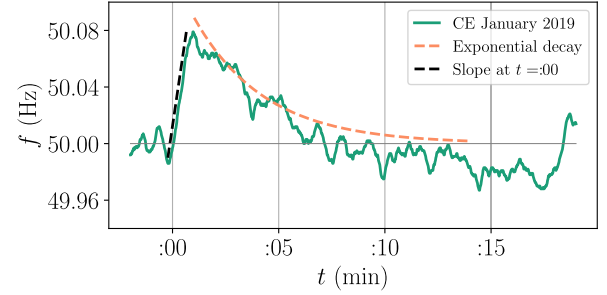


FIGURE 7. Power imbalance ΔP and secondary control c_2 are determined from trading peaks. We investigate the frequency trajectory at a trading peak: The power imbalance ΔP is obtained from the initial slope, i.e., the rate of change of frequency (ROCOF) and the secondary control c_2 from the following exponential decay, see (6). The frequency trajectory is using the CE data from January 10 2019.

The detrending is illustrated in Fig. 5: The same snippet of data from Fig. 1 is shown alongside with the Gaussian kernel detrending. In panel (b) the subtraction of the detrending on the data yields the purely stochastic process governing the power-grid frequency dynamics without deterministic or slow time scale influences. Finally, we extract the first Kramers–Moyal coefficient in panel (c):

$$\mathcal{D}^{(1)}(\omega) = \frac{1}{\Delta t} \langle (\omega(t + \Delta t) - \omega(t)) |_{\omega(t)=\omega} \rangle = -c_1 \omega, \quad (3)$$

where Δt is the sampling rate of the process at hand, which is $\Delta t = 1$ s for our data sets. Furthermore, $\langle \dots |_{\omega(t)=\omega} \rangle$ denotes the following: A spatial average of the difference $(\omega(t + \Delta t) - \omega(t))$ is taken at the point of evaluation $\omega(t) = \omega$, i.e., at a particular frequency ω all differences $(\omega(t + \Delta t) - \omega(t))$ are evaluated and the diffusion $\mathcal{D}^{(1)}$ is obtained as a function of ω . Based on our modelling assumptions, we presume this function to be linear in ω . And, when we apply this to the real data in Fig. 5, we notice that the numerically extracted drift term is indeed well described as a linear function with slope $-c_1$.

3) ESTIMATING THE NOISE AMPLITUDE ϵ

The noise amplitude ϵ is unravelled from data by studying the second Kramers–Moyal coefficient. In our case, we obtain the second conditional moment as

$$\mathcal{D}^{(2)}(\omega) = \frac{1}{\Delta t} \langle (\omega(t + \Delta t) - \omega(t))^2 |_{\omega(t)=\omega} \rangle = \frac{\epsilon^2}{2}, \quad (4)$$

where Δt is again the sampling rate of the process and the empirical $\mathcal{D}^{(2)}(\omega)$ is assumed to approximately constant, based on our model. Computing the second conditional moment $\mathcal{D}^{(2)}$ thereby yields the noise amplitude ϵ . Empirically, we note that the de-trending is not even necessary to determine the correct diffusion coefficient. So we instead compute the diffusion from the original data directly.

We display the diffusion coefficient, i.e., the second Kramers–Moyal coefficient, as a function of the frequency in Fig. 6 for the month of January 2019 for the CE grid. We determine the diffusion coefficient value at 50 Hz and by using (4) thus determine the noise amplitude ϵ .

4) ESTIMATING THE MARKET IMPACT ΔP

To determine both ΔP and c_2 , we have a closer look at the frequency behaviour following a sudden power imbalance. Assuming that the power imbalance is large enough, we can neglect the noise amplitude $\epsilon \approx 0$ as the dynamics close to the power jump are approximately deterministic. Before the power imbalance, we assume that the system is close to the nominal frequency, i.e., $\Delta P = 0$, $\theta \approx 0$ and $\omega \approx 0$. Next, we introduce a power imbalance, e.g. due to trading by setting $\Delta P = P_0$. The equations of motion then are

$$\begin{aligned} \frac{d\theta}{dt} &= \omega, \\ \frac{d\omega}{dt} &= -c_1\omega - c_2\theta + P_0. \end{aligned} \quad (5)$$

A full solution of this driven, damped harmonic oscillator is given by

$$\omega(t) = \frac{P_0 e^{-\frac{1}{2}t(\sqrt{c_1^2 - 4c_2} + c_1)}}{\sqrt{c_1^2 - 4c_2}} \left[e^{t\sqrt{c_1^2 - 4c_2}} - 1 \right]. \quad (6)$$

We evaluate the rate of change of frequency (ROCOF) at the jump time, i.e., at $t = 0$ to be

$$\left. \frac{d\omega}{dt} \right|_{t=0} = P_0, \quad (7)$$

and thereby determine the jump height P_0 , which gives us the power imbalance ΔP , again assuming $\theta(0) = \omega(0) \approx 0$. Recall that we rescaled all variables with the inertia M so that the ROCOF depends on the change of power and the inertia as expected.

Note, while the solution (6) explicitly used the Heaviside function with secondary control (Case A), the ROCOF also determines the power jump in the case of a sawtooth function (Case B). The reason is that the derivative at $t = 0$ is independent of what happens for $t > 0$ and also does not depend on c_1 or c_2 .

5) ESTIMATING THE SECONDARY CONTROL c_2

The estimation of c_2 is only necessary for models that include it, such as Case A with its simple Heaviside function. We know how the trajectory of the angular velocity ω , given by (6), develops following a jump: Initially, the value of ω increases and then decays approximately exponentially back to the reference value.

Since the primary control parameter c_1 is typically much larger than the secondary control parameter c_2 , we make use of the following approximation: $\sqrt{c_1^2 - 4c_2} \approx c_1 - \frac{2c_2}{c_1}$. Thus (6) reduces to

$$\omega(t) = \frac{P_0 e^{-t\frac{c_2}{c_1}}}{c_1 - \frac{2c_2}{c_1}} \left[1 - e^{-t(c_1 - \frac{2c_2}{c_1})} \right]. \quad (8)$$

For larger times $t \gg 1$ s, the second term in (8) decays much faster than the first term. We can therefore further

approximate the angular velocity ω as

$$\omega(t) \sim \exp\left(-\frac{c_2}{c_1}t\right), \quad (9)$$

which allows an estimate of the secondary control c_2 , taken we determined the primary control c_1 earlier. We only need to determine the exponent of the exponential decay, as depicted in Fig. 7. Note that the exponential decay constant does not depend on which trading interval we analyse. For more robust analysis, we perform the fits using the decay following hourly jumps, see also Supplemental Material.

This sequence of parameter estimations allows us to uncover all underlying parameters of the system directly from power-grid frequency measurements. In fact, a single measurement of 60 minutes of data already entails a good ground for estimation but naturally employing as much data as possible yields more reliable parameter estimations, as well as the possibility of error estimation in an efficient way.

IV. CASE STUDY: CONTINENTAL EUROPEAN GRID

With the model properly defined, we now show how it approximates the stochastic behaviour of real frequency trajectories in Europe. The frequency statistics and also market setting differ substantially between different power grids [23]. So, instead of applying each case to all potential power grids, we showcase it on one power grid example where the statistics are well approximated.

Hence, we first apply Case A to data from Continental Europe, Case B to data from Great Britain and finally show that we can also import and utilize real dispatch data to further improve the model predictions in Case C.

A. CASE A: PARAMETER EXTRACTION FOR JANUARY 2019, CENTRAL EUROPE

We analyse power-grid frequency data for the month of January 2019, using measurements provided by the transmission system operator TransnetBW GmbH who operates the German grid in the state Baden-Württemberg [30]. For this month, we estimate the following parameters:

Central Europe, January, 2019.		
ϵ [s^{-2}]	c_1 [s^{-1}]	c_2 [s^{-2}]
0.00105	0.008311	0.000030

For simplicity, we considered the dispatch at the hourly mark as the reference, as can be seen in Fig. 1 to be the strongest driver of the system.

Case A: CE P_0 , January, 2019			
	at :00	at :30	at :15, :45
P_0 [s^{-2}]	0.001641	0.000547	0.000273

Extracting the value, as described, of the ΔP for the hourly mark, we considered the half-hour and quarter-hour trading windows to be 1/3 and 1/6 of the hourly value of ΔP . Notice that there is no limitation in calculating this from data but the results can prove unreliable given the small differences in dispatch. Furthermore, to mimic the structure of the dispatch [24], we take a naïve 6-hour window where

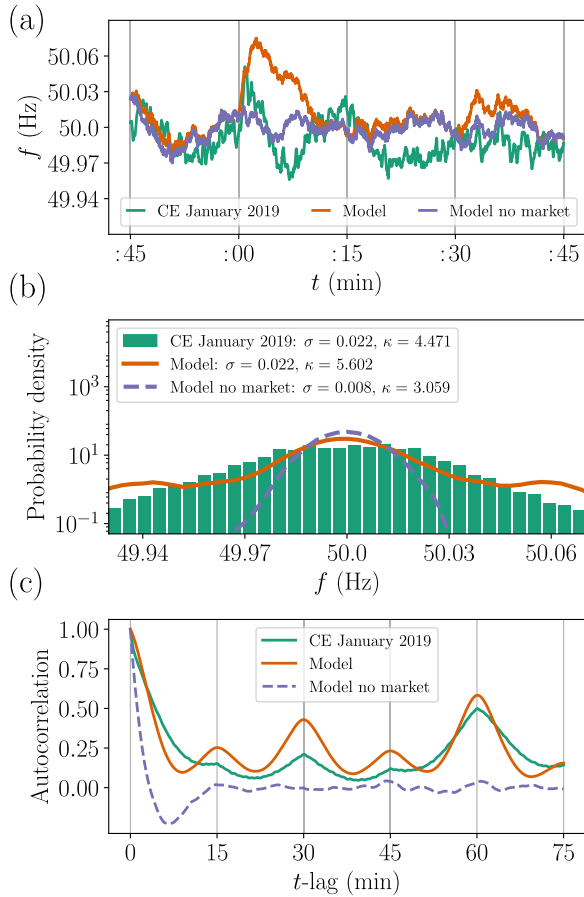


FIGURE 8. Case A: Heaviside dispatch approximates CE trajectories. We compare two days of the power-grid frequency of the Central European (CE) power grid in January 2019 with synthetic data generated by our model (1). For this particular analysis, we utilise Case A that relies on a step function mimicking the jumps of the power mismatch ΔP . The four governing parameters: Noise ϵ , primary c_1 and secondary c_2 control, and power mismatch ΔP parameters are given in Section IV-A, further details are given in the Supplemental Material. (a) We plot a snippet of the power-grid frequency trajectory from the CE data, the model, and the surrogate model without power dispatch. The 15 minute trading intrinsic to the model and the data is highlighted with grey lines. (b) We display the probability density function of the CE data (histogram), the model data (solid line), and the surrogate model data without dispatch (dashed line). Standard deviation and kurtosis of each process are indicated in the legend. (c) We display the autocorrelation of the processes for a time window of 75 minutes, noting the initial exponential decay and regular peaks.

the P_0 jumps are positive values, followed by an equivalent 6-hour window with negative peaks. This should approximate the daily cycles of human daily activity: The work schedule begins: demand increases; Work schedule ends: demand decreases; Private consumption at home begins: demand increases; Night time begins, demand decreases.

Having these parameters at hand, we can now employ our model (1) to integrate synthetic power-grid frequency trajectories. We employ an Euler–Mayurama stochastic integrator, with a time sampling of 0.001 seconds, for a total length of two days, and make use on a step function with changing values every 15 minutes, as formulated in Case A, to mimic the power dispatch curve.

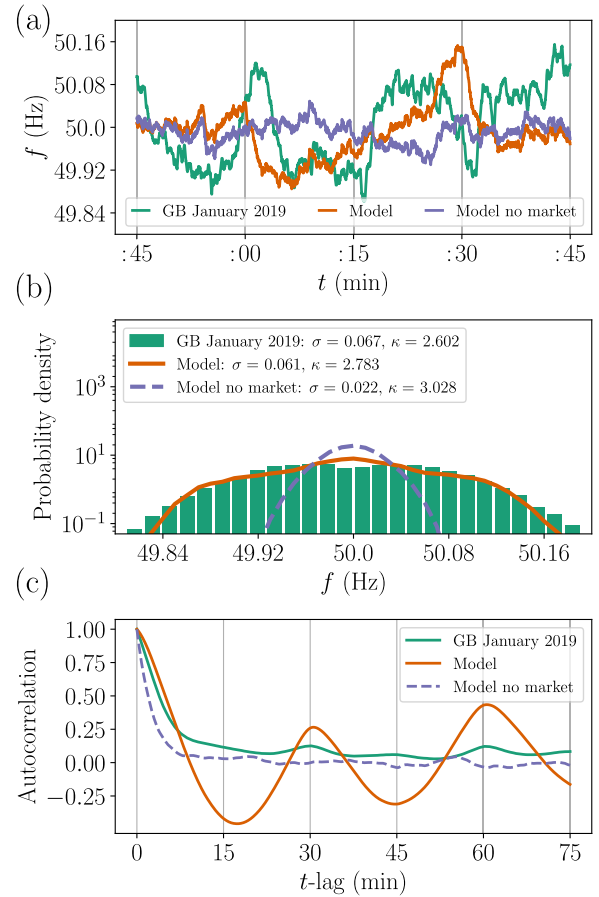


FIGURE 9. Case B: Sawtooth dispatch approximates GB trajectories. We compare two days of the power-grid frequency of the British (GB) power grid in January 2019 with synthetic data generated by our model (1). Here, a sawtooth function is used to describe the mismatch in power ΔP , see Fig. 4, Case B. Noise amplitude ϵ , primary c_1 control, and power mismatch ΔP are given in Section IV-B, see also Supplemental Material for details on parameter estimation. Note that Case B does not use secondary control. (a) We plot snippet of the power-grid frequency trajectory from the GB data, the model, and the surrogate model without power dispatch. (b) We display the probability density function of the GB data (histogram), the model data (solid line), and the surrogate model data without dispatch (dashed line). Standard deviation and kurtosis of each process are indicated in the legend. (c) We display the autocorrelation of the processes for a time window of 75 minutes, noting the initial exponential decay and regular peaks. Contrary to the Heaviside function of Case A, the sawtooth function forces a negative correlation of the system by first driving the system driven to one state and then inverting this trend at the trading interval.

We compare the data, the synthetic model based on (1) and surrogate model without the market structure, i.e., where we set $\Delta P = 0$, in Fig. 8. The introduction of the model without the market allows us to understand concisely the influence of the dispatch on the trajectory of the power-grid frequency, as well as the influence it has on the statistical behaviour of the system.

Several distinct features of the market effect can be seen in Fig. 8: While the surrogate only fluctuates randomly close to the reference frequency, both the real and the synthetic trajectory display surges of the frequency close to the 15 minute trading windows, see panel (a). These large surges lead to a non-Gaussian probability distribution of the power-grid

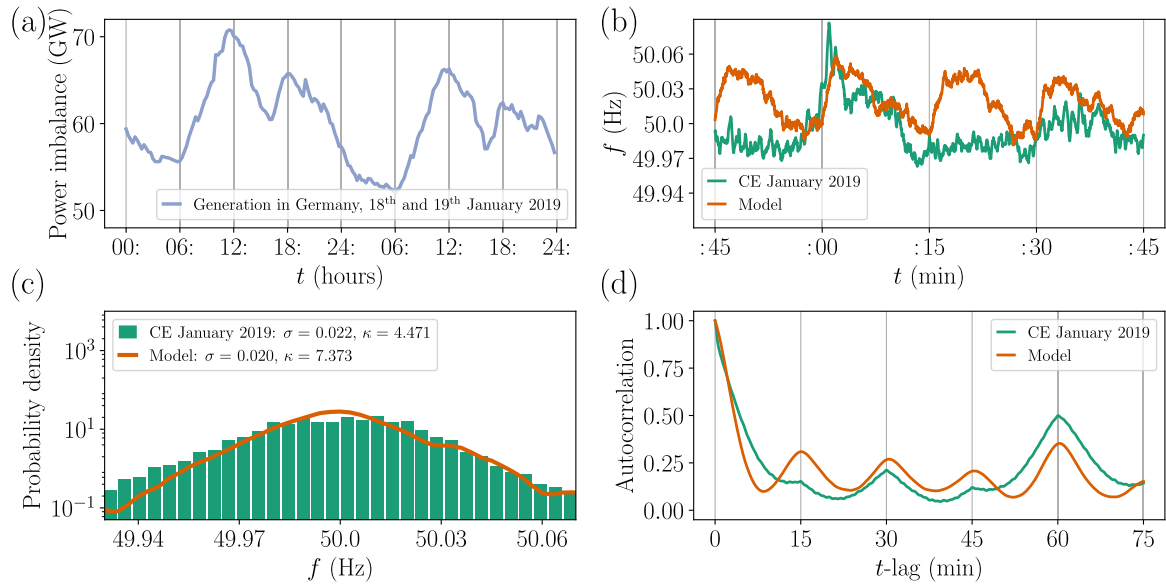


FIGURE 10. Case C: Realistic dispatch trajectories better approximate the real frequency statistics. (a) We use the real dispatch trajectories of the demand in Germany [57] to obtain the correct jumps for the Heaviside function (as in Case A) of the power mismatch ΔP and use our model (1) to generate a synthetic trajectory. (b) The synthetic frequency trajectory statistically resembles the real trajectory for the two day period depicted here. Noise amplitude ϵ , primary c_1 and secondary c_2 control were calculated as described in the Supplemental Material. (c) We display the probability density function of the CE data (histogram), the model data (solid line), and the surrogate model data without dispatch (dashed line). Standard deviation and kurtosis of each process are indicated in the legend. (d) We display the autocorrelation of the processes for a time window of 75 minutes, noting the initial exponential decay and regular peaks.

frequency, evidenced in panel (b). Both the data and the synthetic model with the market display a high kurtosis ($\kappa > 3$), while the surrogate model without any market is essentially Gaussian. This indicates that the market activity has a considerable impact on the distribution of the frequency, specifically its tails. With the market, the system reaches critical values much more often than what would be expected by a normally distributed process. We finally compare the autocorrelation functions of the power-grid frequency for the CE data of January 2019, the modelled data, and the surrogate model in panel (c). We note that the system's scheduled trading/dispatch windows generate defined peaks at exactly 15, 30, 45, and 60 minutes. By comparison, a surrogate system without a market structure displays no correlation peaks at any time lag. Moreover, it is important to notice that *all* peaks in the autocorrelation function are positive valued, both for the synthetic and the real data. This indicates that the system's dispatch is not an uncorrelated random process but the direction of the frequency change is correlated: Frequency surges are more likely followed by more frequency surges and vice versa for frequency sags. The modelled data mimics this with accuracy by implementing an over-simplistic yet successful heuristic argument based on human daily cycles, as explained before.

B. CASE B: PARAMETER EXTRACTION FOR JANUARY 2019, GREAT BRITAIN

Analogously, we analyse data from Great Britain for the month of January 2019, obtained from the British transmission system operator National Grid ESO [60].

Applying the discussed methods, we derive the following parameters

Great Britain, January, 2019			
ϵ [s^{-2}]	ΔP [s^{-2}]	c_1 [s^{-1}]	c_2 [s^{-2}]
0.00205	0.00204	0.00606	#

where in this case we set the value of the secondary control c_2 to be zero. Here, we apply Case B, for two reasons: First, we wish to show that it is also capable of capturing the frequency distributions of a given grid. Second, the British frequency trajectory does not display any clear exponential decay following the trading activity. This is likely caused by a smaller relative trading volume and a larger relative noise amplitude [23]. Both effects also contribute to much smaller autocorrelation peaks at the trading intervals.

We recover the statistics of the British power-grid frequency data with remarkable precision using a sawtooth function for the power dispatch ΔP , see Fig. 9. The GB data exhibits low kurtosis values ($\kappa < 3$), especially when compared to the Continental European values. Our employed model captures the process with high accuracy, when we apply a sawtooth function for ΔP (Case B). Notably, for the case of the surrogate model without market activity, the probability distribution again approximates a Gaussian distribution with kurtosis $\kappa = 3$. Although the autocorrelation function exacerbates the peaks, it captures the initial decay and the trend of regular peaks well. The oversized oscillations arise since we assumed consistent periods of six hours with the same jump and ramp behaviour. In turn, the negative autocorrelation arises as the sawtooth function suddenly changes

the sign of the market effect. Both assumptions are part of a very simple but thereby easy-to-use model of the British grid, which still captures the probability distribution (histogram) very well.

C. CASE C: USING REAL POWER DISPATCH FOR CONTINENTAL EUROPE

Finally, we use real dispatch data from Germany, provided by ENTSO-E [57], to determine the power mismatch ΔP in our model (1) and compare synthetic and real trajectories in Fig. 10. To this end, we simply set the power mismatch ΔP as a Heaviside function based on the real demand for the German grid, i.e., we use the actual demand and assume it stays constant for a given 15 minute interval. While the autocorrelation and the rough shape of the histogram of the model data closely match those of the real data set, we note a substantial difference in the computed kurtosis values. This discrepancy is likely caused by the large variations in the volume of the dispatched power. Here we use data from the German grid to allow a 15 minute resolution. However, the full power dispatch affecting the Continental European grid is given as the sum over all participating countries and would likely be smoother and lead to lower kurtosis values. As noted before, we only require the jump height in ΔP , here as the demand, while the generation enters as the simplified secondary control term $-c_2\theta$. We chose the German data because its time resolution of ΔP is 15 minutes, compared to 1 hour resolution for many other countries. Using such real demand data breaks the symmetrical and regular six hour patterns we have been using so far in Cases A and B. Thereby, we also include larger time scales in the synthetic frequency data since the real demand naturally includes for example daily and weekly cycles. Aside from ΔP , we use the same values as in Case A for the other parameters, i.e., noise ϵ , primary and secondary control c_1 and c_2 . Comparing the synthetic trajectory and derived measures with the real frequency trajectory, we note that including the real demand data improves the approximation further, see Fig. 10. For example the probability density of the real frequency is even better approximated by the synthetic data than in Case A.

V. DISCUSSION

We set out to devise a model to generate realistic synthetic trajectories of the power-grid frequency to be used in simulations of power and control system dynamics and to assist planning and operation of today's and future power grids. To that end, we first showed that the frequency trajectories show both deterministic and stochastic features, leading to non-standard frequency statistics: Heavy tails in the probability distributions and regular autocorrelation peaks pose challenges to properly model the trajectories.

We proposed a simple model combining the deterministic and stochastic aspects of the trajectories. Using stochastic theory and data analysis we were able to extract all essential parameters of the model from real trajectories. We specifically highlighted how the model approximates probability

distributions and autocorrelation functions of realistic grids. A more detailed analysis of the mathematical properties of both real trajectories and the model is presented in [32].

The presented model was designed to be generally applicable, easily extendible and usable, which inevitably requires several simplifications: It does not capture the very short time scale when short-term noise, dynamical behaviour of the rotation machines or switching delays play an important role. Similarly, the model does also not include the long time scale with effects such as synoptic or even seasonal cycles, long-term trading commitment etc. Finally, the model is a stochastic model, i.e., it is not suitable for forecasting of the near future but instead it reproduces critical statistical properties such as large frequency deviations. Conceptually, our modelling approach bridges power engineering, stochastic modelling and data analysis. Power engineering serves as the inspiration to our model building blocks like primary and secondary control. The universality of stochastic modelling is used in formulating the Fokker–Planck equation and deriving both the diffusion coefficient and primary control. Finally, more data analysis tools are necessary to estimate remaining parameters such as the secondary control or the strength of the dispatch or market actions.

Critically, we unveiled how much the market activity influences the tails of the probability distribution, i.e., the probability to observe large deviations from the reference. Comparing models with and without market revealed that just by including the market activity most large events can be explained, consistent with earlier findings [31], [43]. This emphasizes the role the market design has on the stability of the power grid.

The explicit modelling of the market in the stochastic model is specifically interesting when designing new market rules or introducing new business models. As we have seen, the market has a dramatic influence on the stability-defining large deviations. Our model can easily predict the effects on the frequency when shifting from 15-minute to 5-minute dispatch actions or when introducing real-time pricing. New proposals of smart grids, the impact of demand-side management etc. can all be captured by appropriately modifying the power dispatch ΔP of our model. Thereby, we provide guidelines how new concepts and devices can be introduced in the grid without destabilizing it but ideally providing additional stability.

Concluding, our research offers a tool that can be used by natural scientists, mathematicians, engineers, economists or industry practitioners on various questions related to the electricity system. It can be used to plan future grids, such as setting up smart grids and microgrids by providing guidelines on how control parameters should be set to guarantee a certain frequency quality. Executable computer code and easy-to-read pseudo-code of the model and the parameter estimation are provided in the supplementary material.

The model presented here can easily be extended in multiple directions: We could apply more advanced stochastic measures to compare the synthetic trajectory with the real

```

1 # Set of required python libraries
2 import numpy as np
3 from scipy.optimize import curve_fit
4
5 # Library for the gaussian kernel filter
6 from scipy.ndimage.filters import
   gaussian_filter1d
7
8 # Library for calculating Kramers—Moyal
   coefficients
9 from kramersmoyal import km
10
11
12 # Preliminaries
13 # Allocate the power-grid frequency data to
   a numpy array. Make sure the first
14 # entry corresponds to the zero second of
   an hour period, e.g. data[0] is the
15 # start of the data at some HH:00:00
16
17 data = np.loadtxt('location/of/data.txt')
18
19 # if the data is recorded at a reference (e
   .g. 50 Hz), remove the reference
20 data = data - 50.0

```

Listing 1. Load libraries and data.

trajectory, as partially done in [32]. Simultaneously, the frequency dynamics considered here could be extended by voltage amplitude dynamics. While we included primary and secondary control, additional work is necessary if tertiary control should be part of the model. Furthermore, while we only considered constant Gaussian noise, this noise could easily be extended: Either by including explicit non-Gaussian noise [23], as it is observed from wind and solar generators [9] or by making the noise or the control time-dependent, leading to superstatistical modelling [46]. Finally, the proposed model and possible extensions should be systematically compared to alternative power grid frequency forecast methods and their performance versus historic data.

VI. SUPPLEMENTAL MATERIAL

A. PARAMETER EXTRACTION GUIDELINES

Following the mathematical foundations presented in the main text, we present hands-on instructions on how to extract the parameters for the example of a month-long recording of the power-grid frequency in Germany for the month of January 2019. As we focus mainly on specific characteristics of the power dynamics, we calculate, strictly from the data, the noise amplitude ϵ , the power mismatch at the hourly stamp ΔP , the primary and secondary control amplitudes c_1 and c_2 . The procedure follows in a simple manner:

- Noise amplitude ϵ : Utilise the second Kramers—Moyal coefficient, i.e., the diffusion, to extract the noise strength ϵ from the timeseries of the data. Use relation (4) to obtain the value, by taking either the value of the diffusion at $f = 50$ Hz or averaging in windows around $f = 50$ Hz.
- Power mismatch ΔP (for the hourly jumps): Take the first 10 seconds of data just after the hour, e.g. from 12:00:00 to 12:00:10. Calculate the slope of the

```

1 # Noise epsilon
2 # In order to calculate the noise epsilon
   you need to extract the diffusion term
3 # of the stochastic processes. Employ the
   km function from the kramersmoyal
4 # library
5
6 # Retrieve the diffusion coefficient
7 diffusion, space = km(data, powers = [0,
   2], bins = np.array([6000]), bw = 0.05)
8
9 # find the zero frequency
10 zero_frequency = np.argmin(space[0]**2)
11
12 # evaluate the diffusion at that point and
   extract epsilon
13 epsilon = np.sqrt(diffusion[1,
   zero_frequency]*2)

```

Listing 2. Noise ϵ .

frequency increase or decrease in this window with a linear fit. Given that the process displays jumps up and down, i.e., excess and lack of power supply, take the absolute value to obtain the general power mismatch ΔP . Average to obtain the average effect.

- Primary control c_1 : This is a two-step process: Perform a Gaussian kernel de-trending of the data, with a 60-seconds window, to remove the effects of the market and dispatch, so to capture the system's stochastic nature. The choice of a 60-second window ensures one removes only the deterministic characteristics of the frequency trajectory: a smaller window will mimic the noise, a larger window will reflect the overall mean of 50 Hz (60 Hz) of the process. Utilise now the first Kramers—Moyal coefficient, i.e., the drift term, to obtain a negatively tilted line: linearly fit the line around $f = 50$ Hz (or 60 Hz) and extract the slope, which is the drift coefficient of the governing Ornstein—Uhlenbeck process. The slope is the negative primary control $-c_1$.
- Secondary control c_2 : This is the last parameter to calculate, and it depends on the primary control c_1 . Take 900 seconds windows at every hourly jump, similarly to the above calculations for the power mismatch ΔP . Fit (8) to the data snippets (or (6), although strictly mathematically correct, it is harder to fit). Obtain the exponential decay made explicit in (9), i.e., the last term of (8). Input the previously obtained value for the primary control c_1 (step above) to determine the secondary control c_2 .

Having concluded these four steps, we possess all the necessary variables to numerically integrate a synthetic version of the evaluated power-grid frequency.

The simplest and most straightforward method is to implement an Euler—Mayurama integration scheme. This is a scheme identical to a regular Euler integration scheme, incorporating a noise function ξ . This is done by generating a set of normally distributed values with mean $\mu = 0$ and variance $\sigma = \sqrt{\tau}$, with τ the employed time-step of integration. Stochastic integration requires small time-steps, thus we suggest using at least 0.01 seconds, or better even 0.001 seconds.

```

1 # Primary control c1
2 # To calculate the primary control c1 we
  need to employ a two step process.
3 # First remove the general trend by a
  gaussian kernel filtering, then employ
4 # again the km function from the
  kramersmoyal librayr to obtain the
  drift term
5
6 data_filter = gaussian_filter1d(data, sigma
  = 60)
7
8 # Obtain the drift coefficient
9 drift, space = km(data-data_filter, powers
  = [0, 1], bins = np.array([6000]), bw =
  0.01)
10
11 # find the zero frequency
12 mid_point = np.argmin(space[0]**2)
13
14 # Calculate the slope of the drift term,
  which gives the primary control c1.
15 # The fitting is to a line of intercept a
  and slope b. T
16 c1 = curve_fit(lambda t,a,b: a - b*t,
  space[0][mid_point - 500:mid_point +
  500],
17               drift[1,mid_point - 500:mid_point +
  500], p0=(0.0002, 0.005),
18               maxfev=10000
19               )[0][1]

```

Listing 3. Primary control c_1 .

From this store only the 1 second recording to accurately compare with available real power-grid data (if your temporal resolution is different, match it). Other more integrators, such as Runge-Kutta integrators for stochastic equations, can be used to ensure higher precision of the numerical results.

To extract the Kramers–Moyal coefficients there are open source Python ('kramersmoyal') or R ('Langevin') packages, see [61] and [62], respectively.

VII. PSEUDO-CODE

Pseudo-code for extracting the parameters from data, based on the methodology implemented for the Central European power grid. As Supplemental Material, a minimal python code is attached. This was the code used for obtaining the parameters from the data.

In the following we compartmentalise the code in four sections, each corresponding to the parameter recovery of each of the four parameters under analysis: Noise ϵ , primary control c_1 , secondary control c_2 , and dispatch ΔP .

For all cases below, the first step is naturally to import the data

Import data

- Load data

IF data is recorded at 50 hz: data = data - 50

Retrieving the Noise ϵ

- Load module km to obtain Kramers–Moyal coefficients
- diffusion, space = km(data, coefficient = 2)
- find $f=0$ in space
- $\epsilon = \sqrt{\text{diffusion}(\text{space} = 0) \times 2}$

```

1 # Delta P / RoCoF
2 # To calculate the dispatch Delta P
  evaluate the process at every hourly
  jump.
3 # If there is a different dispatch seems,
  change the evaluation to that period.
4 # In principle this can be calculated for
  any interval of power dispatch but
5 # to ensure a good fit, bigger jumps =
  bigger dispatch = better fit
6
7 # Define window of jumps. In this case,
  evaluate the Delta P every hour
8 window = 3600 # 3600 seconds = 1 hour
9
10 # Set the total length to evaluate
11 data_range = data.size // window
12
13 # Initialise an array to record the Delta P
14 Delta_P_slopes = np.zeros(data_range)
15
16 # The jumps are to be evaluate at t=0 but
  since we have noise data, we fit the
17 # first 10 seconds to calculate the slope
18 for j in range(data_range):
19     Delta_P_slopes[j] = curve_fit(lambda t,
20                                   a,b: a + b*t, np.linspace(0,9,10),
21                                   data[3600*(j)
22                                   :3600*(j)+10],
23                                   p0=(0.0, 0.0),
24                                   maxfev=10000
25                                   )[0][1]
26
27 # This results is an array with positive
  and negative slopes, since some
28 # frequency changes are positive (excess
  energy), some are negative. Find the
29 # absolute value for them and take the
  average as the reference Delta P.
30
31 # This is the mean Delta P
32 Delta_P = np.mean(np.abs(Delta_P_slopes))

```

Listing 4. ΔP .

Retrieving the primary control c_1

- Load module km to obtain Kramers–Moyal coefficients
- Load module filter to obtain the Gaussian kernel filtering
- data_filtered = data - filter(data)
- drift, space = km(data_filtered, coefficient = 1)
- find $f=0$ in space
- fit line to drift around space = 0
- $c_1 = -\text{slope of fit}$

Retrieving the dispatch ΔP

FOR every hour:

- fit line to data[first 10 secs]
- save slope to record
- Take absolute of record
- $\Delta P = \text{mean}(\text{abs}(\text{record}))$

Retrieving the secondary control c_2

FOR every hour

- fit curve of (8) to data[900 seconds]
- save exp. decay to record
- $c_2 = \text{mean}(\text{record}) \times c_1$

```

1 # Secondary control c_2
2 # To calculate the secondary control c_2 we
  will need, just as above, snippets
3 # of the hourly jumps and the subsequent
  decay of the frequency back to the
4 # nominal values. Due to the complicated
  frequency behaviour, we will fit an
5 # entire curve to the 900 seconds but we
  shall only extract the decay rate
6
7 # Define window of jumps. In this case,
  evaluate the secondary control c_2
  every
8 # hour
9 window = 3600 # 3600 seconds = 1 hour
10
11 # Set the total length to evaluate
12 data_range = data.size // window
13
14 # Initialise an array to record the Delta P
15 c_2_decays = np.zeros(data_range)
16
17 # Since we have up and down jumps, we have
  to separate the trajectories that
18 # move up and those that move down but we
  still calculate the same decay
19 # behaviour of both
20 for j in range(data_range):
21     # if the frequency trajectory moves
    positively
22     if np.sum((np.diff(data[3600*(j):3600*(
        j)+10]))) > 0:
23         c_2_decays[j] = curve_fit(lambda t,
            a,b,c:
24             a*np.exp(-b*t)*(1-np.exp(-c*t
                +2*b*t)),
25             np.linspace(0,899,900), data
                [3600*(j):3600*(j)+900],
26             p0=(0.08, .0045, 0.035), maxfev
                =10000
27             )[0][1]
28     else:
29         c_2_decays[j] = curve_fit(lambda t,
            a,b,c:
30             -a*np.exp(-b*t)*(1-np.exp(-c*t
                +2*b*t)),
31             np.linspace(0,899,900), data
                [3600*(j):3600*(j)+900],
32             p0=(0.08, .0045, 0.035), maxfev
                =10000
33             )[0][1]
34
35 # We have thus stored the decay rate b of
  every power mismatch in the system.
36 # Due to statistical outliers, discard 20%
  of the data
37
38 # Sort the array and discard 20% of the
  largest values
39 temp_c_2_decays = c_2_decays[np.argsort(
    c_2_decays)][:-c_2_decays.size//5]
40
41 # Recall here that to calculate c_2 you
  need to know c_1
42 c_2 = np.mean(temp_c_2_decays) * c_1

```

Listing 5. Secondary control c_2 .

It is advisable to discard the statistical outliers, since fitting an exponential decay to the frequency data is especially unreliable if the dispatch difference is very small for that period.

VIII. PYTHON MINIMAL-WORKING CODE

Listings 1–5.

ACKNOWLEDGMENT

(Marc Timme and Benjamin Schäfer contributed equally to this work.)

REFERENCES

- [1] J. L. Sawin, F. Sverrisson, J. Rutovitz, S. Dwyer, S. Teske, H. E. Murdock, R. Adib, F. Guerra, L. H. Blanning, and V. Hamirwasia, "Renewables 2018-global status report," REN21-Renewables Now, Paris, France, Tech. Rep. REN21, 2019. [Online]. Available: https://www.ren21.net/wp-content/uploads/2019/05/gsr_2019_full_report_en.pdf
- [2] J. Rodríguez-Molina, M. Martínez-Núñez, J.-F. Martínez, and W. Pérez-Aguilar, "Business models in the smart grid: Challenges, opportunities and proposals for prosumer profitability," *Energies*, vol. 7, no. 9, pp. 6142–6171, Sep. 2014.
- [3] S. Schultz. (2019). *Deutsche Netzbetreiber Kämpfen Mit Akuter Stromnot*. SPIEGEL ONLINE. [Online]. Available: <https://www.spiegel.de/wirtschaft/unternehmen/stromnetz-deutsche-netzbetreiber-kaempften-mit-akuter-stromnot-a-1275323.html>,
- [4] (2020). *Energiewende-Index*, McKinsey & Company. [Online]. Available: <https://www.mckinsey.de/branchen/chemie-energie-rohstoffe/energiewende-index>
- [5] N. L. Panwar, S. C. Kaushik, and S. Kothari, "Role of renewable energy sources in environmental protection: A review," *Renew. Sustain. Energy Rev.*, vol. 15, no. 3, pp. 1513–1524, Apr. 2011.
- [6] M. L. Tuballa and M. L. Abundo, "A review of the development of smart grid technologies," *Renew. Sustain. Energy Rev.*, vol. 59, pp. 710–725, Jun. 2016.
- [7] B. H. Obama, "Presidential policy directive 21: Critical infrastructure security and resilience," White House, Office Press Secretary, Washington, DC, USA, Tech. Rep., 2013.
- [8] P. Kundur, N. J. Balu, and M. G. Lauby, *Power System Stability and Control*, vol. 7. New York, NY, USA: McGraw-Hill, 1994.
- [9] M. Anvari, G. Lohmann, M. Wächter, P. Milan, E. Lorenz, D. Heinemann, M. R. R. Tabar, and J. Peinke, "Short term fluctuations of wind and solar power systems," *New J. Phys.*, vol. 18, no. 6, Jun. 2016, Art. no. 063027.
- [10] M. F. Wolff, K. Schmietendorf, P. G. Lind, O. Kamps, J. Peinke, and P. Maass, "Heterogeneities in electricity grids strongly enhance non-Gaussian features of frequency fluctuations under stochastic power input," 2019, *arXiv:1908.07997*. [Online]. Available: <http://arxiv.org/abs/1908.07997>
- [11] M. Rohden, A. Sorge, M. Timme, and D. Witthaut, "Self-organized synchronization in decentralized power grids," *Phys. Rev. Lett.*, vol. 109, no. 6, Aug. 2012, Art. no. 064101.
- [12] T. Walter, *Smart Grid neu gedacht: Ein Lösungsvorschlag zur Diskussion in VDE/ETG*. Frankfurt, Germany: VDE Verband der Elektrotechnik, 2014.
- [13] B. Schäfer, M. Matthiae, M. Timme, and D. Witthaut, "Decentral smart grid control," *New J. Phys.*, vol. 17, no. 1, Jan. 2015, Art. no. 015002.
- [14] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—The new and improved power grid: A survey," *IEEE Commun. Surveys Tut.*, vol. 14, no. 4, pp. 944–980, 4th Quart., 2012.
- [15] J. Weber, M. Reyers, C. Beck, M. Timme, J. G. Pinto, D. Witthaut, and B. Schäfer, "Wind power persistence characterized by superstatistics," 2018, *arXiv:1810.06391*. [Online]. Available: <http://arxiv.org/abs/1810.06391>
- [16] G. Filatrella, A. H. Nielsen, and N. F. Pedersen, "Analysis of a power grid using a Kuramoto-like model," *Eur. Phys. J. B*, vol. 61, no. 4, pp. 485–491, Feb. 2008.
- [17] T. Nishikawa and A. E. Motter, "Comparative analysis of existing models for power-grid synchronization," *New J. Phys.*, vol. 17, no. 1, Jan. 2015, Art. no. 015012.
- [18] K. Schmietendorf, J. Peinke, and O. Kamps, "The impact of turbulent renewable energy production on power grid stability and quality," *Eur. Phys. J. B*, vol. 90, no. 11, Nov. 2017.
- [19] P. C. Böttcher, A. Otto, S. Kettemann, and C. Agert, "Time delay effects in the control of synchronous electricity grids," 2019, *arXiv:1907.13370*. [Online]. Available: <http://arxiv.org/abs/1907.13370>

- [20] H. Haehne, J. Schottler, M. Waechter, J. Peinke, and O. Kamps, "The footprint of atmospheric turbulence in power grid frequency measurements," *Europhys. Lett.*, vol. 121, no. 3, Feb. 2018, Art. no. 30001.
- [21] H. Zhang and P. Li, "Probabilistic analysis for optimal power flow under uncertainty," *IET Gener. Transm. Distrib.*, vol. 4, no. 5, p. 553, 2010.
- [22] B. Schäfer, M. Matthiae, X. Zhang, M. Rohden, M. Timme, and D. Witthaut, "Escape routes, weak links, and desynchronization in fluctuation-driven networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, no. 6, Jun. 2017.
- [23] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme, "Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics," *Nature Energy*, vol. 3, pp. 119–126, Jan. 2018.
- [24] T. Weissbach and E. Welfonder, "High frequency deviations within the European Power System: Origins and proposals for improvement," in *Proc. IEEE/PES Power Syst. Conf. Expo.*, Mar. 2009, pp. 1–6.
- [25] J. Scoltock, T. Geyer, and U. K. Madawala, "Model predictive direct power control for grid-connected NPC converters," *IEEE Trans. Ind. Electron.*, vol. 62, no. 9, pp. 5319–5328, Sep. 2015.
- [26] D. Dong, J. Li, D. Boroyevich, P. Mattavelli, I. Cvetkovic, and Y. Xue, "Frequency behavior and its stability of grid-interface converter in distributed generation systems," in *Proc. 27th Annu. IEEE Appl. Power Electron. Conf. Expo. (APEC)*, Feb. 2012, pp. 1887–1893.
- [27] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, Sep. 2007.
- [28] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Oct. 2011, pp. 528–533.
- [29] E. B. T. Tchuiseu, D. Gomila, D. Brunner, and P. Colet, "Effects of dynamic-demand-control appliances on the power grid frequency," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 96, no. 2, Aug. 2017, Art. no. 022302.
- [30] *Regelenergie Bedarf + Abruf*, TransnetBW GmbH, Stuttgart, Germany, 2019.
- [31] B. Schafer, M. Timme, and D. Witthaut, "Isolating the impact of trading on grid frequency fluctuations," in *Proc. IEEE PES Innov. Smart Grid technol. Conf. Eur. (ISGT-Europe)*, Oct. 2018, pp. 1–5.
- [32] M. Anvari, L. R. G. ao, M. Timme, B. Schäfer, D. Witthaut, and H. Kantz, "Stochastic properties of the frequency dynamics in real and synthetic power grids," *Phys. Rev. Res.*, to be published.
- [33] C. W. Gardiner, *Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences*. Berlin, Germany: Springer-Verlag, 1985.
- [34] J. Machowski, J. Bialek, and J. Bumby, *Power System Dynamics: Stability and Control*. Hoboken, NJ, USA: Wiley, 2011.
- [35] A. Oudalov, D. Chartouni, and C. Ohler, "Optimizing a battery energy storage system for primary frequency control," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1259–1266, Aug. 2007.
- [36] J. Allen Wood and F. Bruce Wollenberg, *Power Generation, Operation, and Control*. Hoboken, NJ, USA: Wiley, 2012.
- [37] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbič, "Foundations and challenges of low-inertia systems," in *Proc. Power Syst. Comput. Conf. (PSCC)*, Jun. 2018, pp. 1–25.
- [38] H.-P. Beck and R. Hesse, "Virtual synchronous machine," in *Proc. 9th Int. Conf. Electr. Power Quality Utilisation*, Oct. 2007, pp. 1–6.
- [39] P. Kundur, "Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1387–1401, May 2004.
- [40] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE Trans. Ind. Informat.*, vol. 7, no. 3, pp. 381–388, Aug. 2011.
- [41] M. R. Kovacevic, C. G. Pflug, and T. M. Vespucchi, *Handbook of Risk Management in Energy Production and Trading*, vol. 199. New York, NY, USA: Springer, 2013.
- [42] National Academies of Sciences Engineering and Medicine, *The Power of Change: Innovation for Development and Deployment of Increasingly Clean Electric Power Technologies*. Washington, DC, USA: National Academies Press, 2016.
- [43] T. Weißbach and E. Welfonder, "High frequency deviations within the European power system: Origins and proposals for improvement," *VGB powertech*, vol. 89, no. 6, p. 26, 2009.
- [44] E. González-Romera, M. A. Jaramillo-Morán, and D. Carmona-Fernández, "Forecasting of the electric energy demand trend and monthly fluctuation with neural networks," *Comput. Ind. Eng.*, vol. 52, no. 3, pp. 336–343, Apr. 2007.
- [45] P. Milan, M. Wächter, and J. Peinke, "Turbulent character of wind energy," *Phys. Rev. Lett.*, vol. 110, no. 13, Mar. 2013, Art. no. 138701.
- [46] C. Beck and E. G. D. Cohen, "Superstatistics," *Phys. A, Stat. Mech. Appl.*, vol. 322, pp. 267–275, May 2003.
- [47] P. Milan, M. Wächter, and J. Peinke, "Stochastic modeling and performance monitoring of wind farm power production," *J. Renew. Sustain. Energy*, vol. 6, no. 3, May 2014, Art. no. 033119.
- [48] C. Beck, E. G. D. Cohen, and H. L. Swinney, "From time series to superstatistics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 72, no. 5, Nov. 2005, Art. no. 056133.
- [49] A. Ulbig, T. S. Borsche, and G. Andersson, "Impact of low rotational inertia on power system stability and operation," *IFAC Proc. Volumes*, vol. 47, no. 3, pp. 7290–7297, 2014.
- [50] E. Weitenberg, Y. Jiang, C. Zhao, E. Mallada, C. De Persis, and F. Dörfler, "Robust decentralized secondary frequency control in power systems: Merits and trade-offs," 2017, *arXiv:1711.07332*. [Online]. Available: <http://arxiv.org/abs/1711.07332>
- [51] E. B. T. Tchuiseu, D. Gomila, P. Colet, D. Witthaut, M. Timme, and B. Schäfer, "Curing Braess' paradox by secondary control in power grids," *New J. Phys.*, vol. 20, no. 8, 2018, Art. no. 083005.
- [52] C. Wang, C. Grebogi, and M. S. Baptista, "Control and prediction for blackouts caused by frequency collapse in smart grids," *Chaos*, vol. 26, no. 9, Sep. 2016, Art. no. 093119.
- [53] W. John Simpson-Porco, F. Dörfler, and F. Bullo, "Droop-controlled inverters are kuramoto oscillators," *IFAC Proc. Volumes*, vol. 45, no. 26, pp. 264–269, 2012.
- [54] A. T. Hammid, M. Hojabri, M. H. Sulaiman, A. N. Abdalla, and A. A. Kadhim, "Load frequency control for hydropower plants using pid controller," *J. Telecommun., Electron. Comput. Eng.*, vol. 8, no. 10, pp. 47–51, 2016.
- [55] E. D. Dongmo, P. Colet, and P. Wofo, "Power grid enhanced resilience using proportional and derivative control with delayed feedback," *Eur. Phys. J. B*, vol. 90, no. 1, p. 6, Jan. 2017.
- [56] R. Martyr, B. Schäfer, C. Beck, and V. Latora, "Benchmarking the performance of controllers for power grid transient stability," *Sustain. Energy, Grids Netw.*, vol. 18, Jun. 2019, Art. no. 100215.
- [57] ENTSO-E. (2019). *Generation Forecast-Day Ahead*. [Online]. Available: <https://transparency.entsoe.eu/generation/r2/dayAheadAggregatedGeneration/show>
- [58] H. Risken, *The Fokker-Planck Equation*. Berlin, Germany: Springer, 1984.
- [59] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, "Approaching complexity by stochastic methods: From biological systems to turbulence," *Phys. Rep.*, vol. 506, no. 5, pp. 87–162, Sep. 2011.
- [60] *Historic Frequency Data*, National Grid ESO, Warwick, U.K., 2019. [Online]. Available: <https://www.nationalgrideso.com/balancing-services/frequency-response-services/historic-frequency-data>
- [61] L. Gorjão and F. Meirinhos, "kramersmoyal: Kramers–Moyal coefficients for stochastic processes," *J. Open Source Softw.*, vol. 4, no. 44, p. 1693, Dec. 2019.
- [62] P. Rinn, P. G. Lind, M. Wächter, and J. Peinke, "The Langevin approach: An R package for modeling Markov processes," *J. Open Res. Softw.*, vol. 4, no. 1, pp. 1–19, Aug. 2016.



LEONARDO RYDIN GORJÃO received the B.Sc. degree in physics from the University of Lisbon, Portugal, and the M.Sc. degree in physics from the University of Bonn, Germany. He is currently pursuing the Ph.D. degree with the University of Cologne and Forschungszentrum Jülich, Germany.



Impact Research, Power Grid Group, Potsdam, Germany.

MEHRNAZ ANVARI received the M.Sc. degree in physics from the Iran University of Science and Technology, Tehran, Iran, in 2010, and the Ph.D. degree from the ForWind Institute, University of Oldenburg, Germany, in 2016, where she was awarded the George-Christoph-Lichtenberg Scholarship for this period. She held a Postdoctoral position at the Max Planck Institute for Physics of Complex Systems. She is currently a Postdoctoral Researcher with the Potsdam Institute for Climate



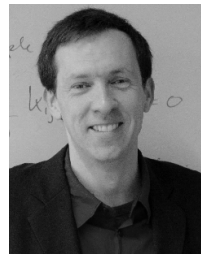
ing a Research Group at Forschungszentrum Jülich, Germany. He is currently an Assistant Professor with the University of Cologne.

DIRK WITTHAUT received the M.Sc. and Ph.D. degrees in physics from the Technical University of Kaiserslautern, Kaiserslautern, Germany, in 2004 and 2007, respectively. He has worked as a Postdoctoral Researcher at the Niels Bohr Institute, Copenhagen, Denmark, and at the Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany. He has been a Guest Lecturer with the Kigali Institute for Science and Technology, Rwanda. Since 2014, he is lead-



Physics of Complex Systems, Dresden, and also an Adjunct Professor of statistical physics with the Institute for Theoretical Physics, Technical University of Dresden.

HOLGER KANTZ received the Diploma degree in physics from the University of Wuppertal, in 1986, and the Ph.D. degree in physics, under the supervision of Peter Grassberger, in 1989. After a period as a Postdoctoral Fellow in Florence, he returned to Wuppertal, in 1991, as a Scientific and Teaching Assistant. He completed his Habilitation in theoretical physics, in 1996. He is the Head of the Nonlinear Dynamics and Time Series Analysis Research Group, Max Planck Institute for the



visiting faculty at ETH Zurich. He is currently a Strategic Professor and the Head of the Chair for Network Dynamics at the Cluster of Excellence Center for Advancing Electronics Dresden (cfaed) and the Institute for Theoretical Physics, TU Dresden. He is also the Co-Chair of the Division of Socio-Economic Physics of the German Physical Society (DPG). Since 2018, he has been an Honorary Member of Lakeside Labs, Klagenfurt. His research integrates first principles theory with data-driven modeling to establish generic fundamental insights that drive applications of complex dynamical systems, including bio-inspired information processing, energy systems, collective mobility and transport, as well as systemic sustainability.

MARC TIMME studied physics and mathematics in Würzburg, Stony Brook (USA), and Göttingen. After working as a Postdoctoral Researcher at the Max Planck Institute for Flow Research and as a Research Scholar at Cornell University, Ithaca, NY, USA, he was selected to head a broadly cross-disciplinary Max Planck Research Group on Network Dynamics at the Max Planck Institute for Dynamics and Self-Organization. He held a Visiting Professorship at TU Darmstadt and was a



currently the Chair of the Statistical and Nonlinear Physics Division of the European Physical Society (EPS). His research interests include mathematical and stochastic modeling aspects of complex systems and applications to real-world systems.

CHRISTIAN BECK received the M.Sc. and Ph.D. degrees in theoretical physics from RWTH Aachen, Germany. After some years abroad in Warwick, Budapest, Copenhagen, and the University of Maryland, he settled in London, where he is currently a Professor of applied mathematics with the Queen Mary University of London, and the Head of the Dynamical Systems and Statistical Physics Group. He is also a Fellow of the Institute of Mathematics and its Applications (FIMA) and



he has been working as a Marie Skłodowska-Curie Research Fellow at the Queen Mary University of London.

BENJAMIN SCHÄFER received the Diploma degree in physics from the University of Magdeburg, Germany, in 2013. Pursuing his Ph.D. in Göttingen, Germany, London, U.K., and Tokyo, Japan, he received the Ph.D. degree in physics from the University of Göttingen, in 2017. He has worked as a Postdoctoral Researcher at the Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany, and the Technical University Dresden, Germany. Since 2019,

...

2.1.2 Publication #2

M. Anvari, L. Rydin Gorjão, M. Timme, D. Witthaut, B. Schäfer, and H. Kantz. *Stochastic properties of the frequency dynamics in real and synthetic power grids*. Physical Review Research **2**(1), 2020, p. 013339. Ref. [2].

Status: published

Stochastic properties of the frequency dynamics in real and synthetic power grids

Mehrnaz Anvari^{1,2,*}, Leonardo Rydin Gorjão^{3,4,†}, Marc Timme^{5,6,‡}, Dirk Witthaut^{3,4,§},
Benjamin Schäfer^{7,5,6,||} and Holger Kantz¹

¹Max Planck Institute for the Physics of Complex Systems (MPIPKS), 01187 Dresden, Germany

²Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, D-14412 Potsdam, Germany

³Forschungszentrum Jülich, Institute for Energy and Climate Research - Systems Analysis and Technology Evaluation (IEK-STE),
52428 Jülich, Germany

⁴Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany

⁵Chair for Network Dynamics, Center for Advancing Electronics Dresden (cfaed) and Institute for Theoretical Physics,
Technical University of Dresden, 01062 Dresden, Germany

⁶Network Dynamics, Max Planck Institute for Dynamics and Self-Organization (MPIDS), 37077 Göttingen, Germany

⁷School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom



(Received 1 October 2019; accepted 19 December 2019; published 19 March 2020)

The frequency constitutes a key state variable of electrical power grids. However, as the frequency is subject to several sources of fluctuations, ranging from renewable volatility to demand fluctuations and dispatch, it is strongly dynamic. Yet, the statistical and stochastic properties of the frequency fluctuation dynamics are far from fully understood. Here we analyze properties of power-grid frequency trajectories recorded from different synchronous regions. We highlight the non-Gaussian and still approximately Markovian nature of the frequency statistics. Furthermore, we find that the frequency displays significant fluctuations exactly at the time intervals of regulation and trading, confirming the need of having a regulatory and market design that respects the technical and dynamical constraints in future highly renewable power grids. Finally, employing a recently proposed synthetic model for the frequency dynamics, we combine our statistical and stochastic analysis and analyze in how far dynamically modeled frequency properties match the ones of real trajectories.

DOI: [10.1103/PhysRevResearch.2.013339](https://doi.org/10.1103/PhysRevResearch.2.013339)

I. INTRODUCTION

A stable electric power supply is essential for the functioning of our society [1]. The ongoing energy transition towards renewable generation fundamentally changes the conditions for the operation of the power system [2]. A better understanding of the dynamics, control, and variability of this highly complex system is needed to ensure stability in a rapidly changing environment [3,4].

The power-grid frequency is the central observable for the control of AC electric power grids, as it directly reflects the balance of the grid: A surplus of feed-in power increases the frequency and a shortage reduces the frequency [5]. Observing the frequency of the power grid can thus provide deep insights into the dynamical stability of the grid as well as the operation of the control system and the economic dispatch of generators. In today's system strict operational boundaries are

imposed on the frequency and the rate of change of frequency [6]. For example, in the Central European power grid (CE), the stable operational boundary for frequency variations is set at ± 200 Hz. Moreover, if the frequency deviates more than $\Delta f = \pm 20$ Hz, the existing control systems, i.e., primary and secondary control, are activated to compensate the imbalance in the power grid and to return the frequency to the nominal one [7].

These control mechanisms and operational boundaries are especially interesting when designing new grids involving concepts such as *smart grids* [8], *prosumers* [9], or *microgrids* [10], and their interaction with the grid frequency. Furthermore, due to the increased usage of renewable energies, synchronous machines are replaced by power electronics, such as inverters, posing additional challenges on ensuring frequency stability [11]. Inverter-based generators do not have any innate inertia, leading to the frequency of the power grid becoming more volatile, unless additional stabilizers are included in the system [12].

A more sophisticated analysis of the power-grid frequency dynamics is paramount, as all power generators and consumers have to ensure the stability of the grid in the presence of many effects simultaneously impinging on it. In such analyses it is both relevant to study existing power grids [13] as well as to evaluate any forecasts and models of the frequency dynamics expected in future grids [14].

Despite the strict operational boundaries for frequency variations, numerous different sources of disturbances

*anvari@pik-potsdam.de

†l.rydin.gorjao@fz-juelich.de

‡marc.timme@tu-dresden.de

§d.witthaut@fz-juelich.de

||b.schaefer@qmul.ac.uk

introduce measurable variations of the frequency over time. Important sources introducing fluctuations to the grid frequency include consumers, renewable energies, and the dispatch of power plants via the energy market. Recent research shows that today's demand fluctuations contribute substantially to uncertainties in the power balance [15–17]. Moreover, intermittent renewable energies influence the frequency first due to their stochastic and often non-Gaussian power feed-in [18,19], and second due to the decreasing the inertia in the power grid, as mentioned above. Hence, to operate energy systems with a high share of renewable energies, a solid understanding of the impact of fluctuating feed-in on the grid's frequency is necessary. Previous studies described the stochastic behavior of the grid frequency using stochastic optimization [20], a simulated robustness analysis [21], Fokker-Planck approaches [22,23], or tracing the impact of wind feed-in on the grid frequency [24,25], and the integration of storage systems to improve the frequency quality in the presence of wind power [26]. However, the mathematical properties of the underlying stochastic process have not been studied comprehensively.

In addition to the aforementioned stochastic disturbances, trading affects the grid frequency by scheduled deterministic periodic events, e.g., dispatch actions on the energy market cause brief jumps of the frequency [13,23,27]. While deterministic disturbances have been observed for various grids [13,28], no comprehensive model exists to describe the market interaction with the grid frequency quantitatively. We thus aim for a dynamical model of the power-grid frequency including the role of trading and regulator action in the power grid. Such a model may help especially to plan future grids with a high share of renewable energies. Volatile renewable energies, such as wind and solar power, are unpredictable and thus cannot be used to balance the grid frequency following trading actions. Instead, it is fundamental to understand the interplay between the stochastic dynamics of unpredictable fluctuations and the deterministic characteristics of the energy market.

Here we first review essential statistical properties and the temporal evolution of the frequency of real-world power grids. With a special focus on the deterministic fluctuations at trading and dispatch times. Our approach provides a method to obtain bountiful information on the power-grid frequency that can be obtained from simple measurements. Next we introduce our stochastic model to regenerate the frequency dynamics and explain how we estimate its parameters solely from the power-grid trajectory. Finally, we demonstrate how our model reproduces key aspects of the stochastic and deterministic behavior of real trajectories.

II. POWER-GRID FREQUENCY OVERVIEW

The power-grid frequency displays several characteristic features, such as non-Gaussian distributions, an exponential decay of the autocorrelation, and regular impacts by trading [23]. We extend earlier studies by uncovering other stochastic properties of power-grid frequency, namely addressing the questions of Markovianity, linearity, and stationarity of the data. Specifically, we investigate the recorded frequency from Great Britain (GB) [29], and from two different regions in central Europe (CE). The two data samples of CE have

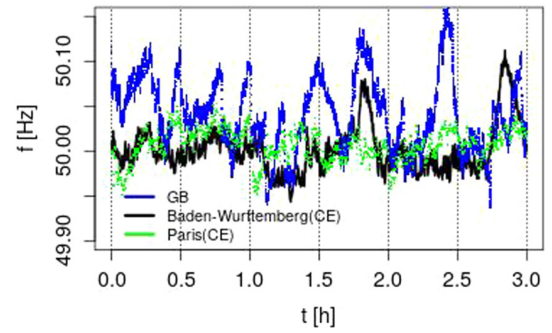


FIG. 1. The power-grid frequency fluctuates over time, with differences between distinct regions. Displayed are 3 h of frequency trajectories on March 1st for Paris, Baden-Württemberg (both CE), and GB. The data sets belong to 2015, 2016, and 2017, respectively, for Paris, GB, and Baden-Württemberg. Note that the Baden-Württemberg and Paris data are from different years, while still displaying similar statistics.

been recorded in Paris (France) [30] and Baden-Württemberg (southwest of Germany) [31]. The time resolution of data sets are 1, 10, and 1 s, respectively for GB, Paris, and Baden-Württemberg. We analyze data spanning over one year: 2015 for France, 2016 for GB, and 2017 for Baden-Württemberg. The final section addresses the modeling following the data from Baden-Württemberg. A direct observation of the frequency of the three samples (Great Britain, Paris, Baden-Württemberg) during three arbitrarily chosen hours in March reveals substantial differences in the fluctuation patterns, see Fig. 1. The range of variations in GB is larger than in the other two frequency data sets. The reason being, the primary control in GB is only activated for frequency deviations of at least ± 200 Hz, while the other frequency sets belong to the CE grid, where control is activated at ± 20 Hz. Consequently the CE data set has smaller overall fluctuations and a lower standard deviation.

In contrast to many random processes, the values of the power-grid frequencies do not strictly follow Gaussian (normal) distributions [32,33]. Instead, the distributions display heavy tails, where large deviations occur much more frequently than anticipated from a normal distribution. In Fig. 2 the frequency and increment frequency distributions of GB and Baden-Württemberg are shown. As both Paris and Baden-Württemberg belong to the CE power grid, they have similar (but not identical) statistical properties. Therefore, for the rest of this section, we focus our analysis on the frequency measurements from Baden-Württemberg as an example, and where we aim to refer to general statistic features, we refer to the CE grid. Comparing the frequency probability distribution function (PDF) with the best-fitting normal distribution, highlights the non-Gaussian properties of the frequency PDF of CE, which has a kurtosis 4.23, Fig. 2(c). The kurtosis, the normalized fourth moment, measures the heavy-tailedness of a distribution, see, e.g., [34]. Any value of the kurtosis larger than the that of a normal distribution ($\kappa_{\text{normal}} = 3$) indicates heavy tails [35]. The frequency distribution for GB breaks the symmetry expected from a normal distribution and exhibits a skewness of 0.191, see Fig. 2(a). The skewness, the normalized third moment β , measures how skewed, i.e.,

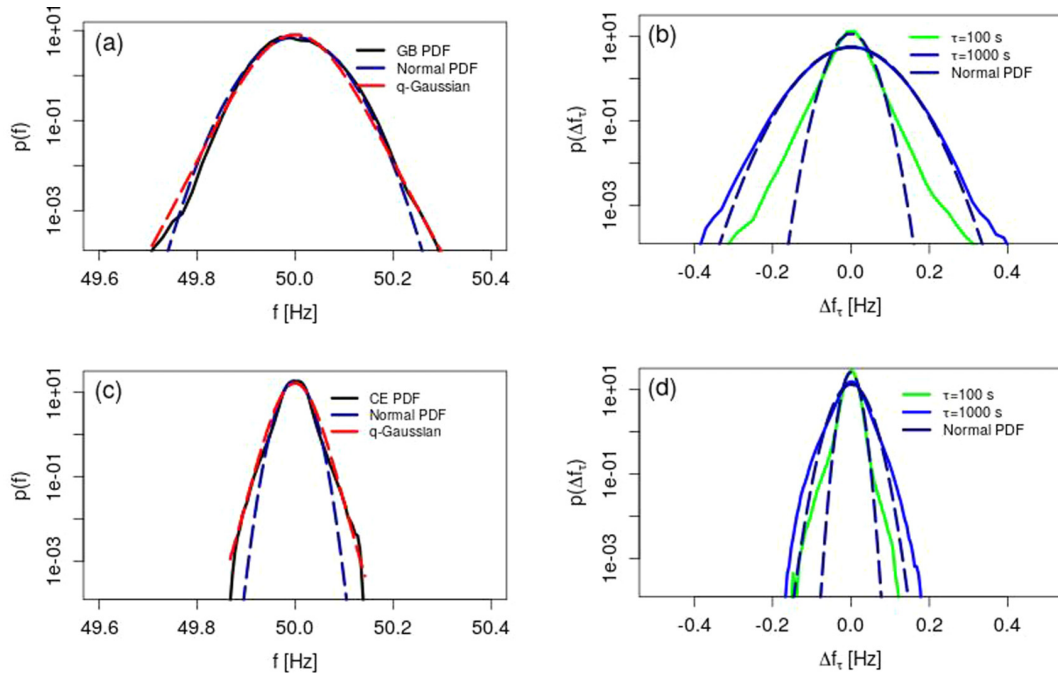


FIG. 2. Both the PDFs of the frequency and of the frequency increments display non-Gaussian features. We compare the PDF of the frequency with the most-likely Gaussian fit (blue curve) and q-Gaussian (red curve), for (a) the GB grid and (c) the CE grid evaluating the Baden-Württemberg time series. We observe an asymmetry (nonzero skewness β) in the GB data with the deformation parameter $q = 0.95$ and pronounced heavy tails (high kurtosis κ) in the CE data with $q = 1.1$. Increment statistics in (b) GB and (d) CE grid were carried out for different time lags. Short-time lag ($\tau = 100$ s) displays more pronounced deviations from Gaussianity (dashed lines) than larger time lags.

asymmetric, a distribution is. For a normal distribution, the skewness is zero. Furthermore, based on the shape of the PDFs, large deviations of the power-grid frequency towards very low frequencies occur more often in the GB grid, while deviations towards higher frequencies are more common in the CE grid. We note that both skewness and kurtosis statistics depend on the sample size, but the observed non-Gaussian features are genuine since we do use large data sets with high sample frequency. Instead of normal distributions, the observed statistics is possibly better described by Lévy-stable or q-Gaussian distributions [23].

The frequency increment statistics also display non-Gaussian features. We estimate the probability to observe large fluctuations on short timescales by computing frequency increments, i.e., $\Delta f_\tau = f(t + \tau) - f(t)$, see Figs. 2(b) and 2(d), for $\tau = 100$ s and $\tau = 1000$ s, respectively. Next, we compare the observed increment probabilities with the best Gaussian fit: Frequency variations of the order of 210 mHz within 100 mHz occur in the GB frequency data set 10^5 times more often than expected for Gaussian processes. For the Baden-Württemberg data, frequency variations ~ 60 mHz occur 100 times more often compared to a Gaussian distribution. The increment frequency statistics indicates that the frequency on the short timescale is particularly subject to large fluctuations. Potentially new control systems or market mechanisms are necessary to compensate the power imbalance in the power grid on short timescales. In contrast, the shape of the frequency and frequency increment PDF become similar for larger time lags, such as $\tau = 1000$, and the deviation from Gaussianity is not as extreme as for the short timescale, see Figs. 2(b) and 2(d).

To obtain more information from the frequency trajectory, we investigate the autocorrelation and its decay for the frequency data sets. The autocorrelation measures the correlation of a signal with itself at a later time. High correlation values indicate that a large signal is typically followed by still a large signal and vice versa. The power-grid frequency autocorrelation decays approximately exponentially as a function of the time lag Δt for short-time lags, see [23] and Fig. 3. Several prototypical stochastic processes, such as the Ornstein-Uhlenbeck process, display a similar decay, following precisely an exponential function [36]

$$c(\Delta t) = \langle f(t)f(t + \tau) \rangle, \quad (1)$$

$$c^{\text{OU}}(\Delta t) = \exp(-\alpha \Delta t), \quad (2)$$

with a damping constant α . While initially the system is highly correlated with its own history, this damping will cause a decorrelation. Naturally, distinct power grids will have their specific characteristic damping constant. A least squares fit of an exponential decay (2) to the data yields α^{-1} which is ~ 385 s for the GB grid and ~ 312 s for the CE grid respectively, see Fig. 3(a).

Another feature of the autocorrelation are the regular peaks every 15 min, which are highlighted with black arrows in Fig. 3. These peaks are caused by a mismatch of power supply and demand [13,27,32]. In most electricity grids the operation of dispatchable power plants is scheduled in 1 h blocks, where additional (shorter) 30 and 15 min intervals might exist. Hence the generation curve is steplike, while the demand varies continuously. From step to step, the power

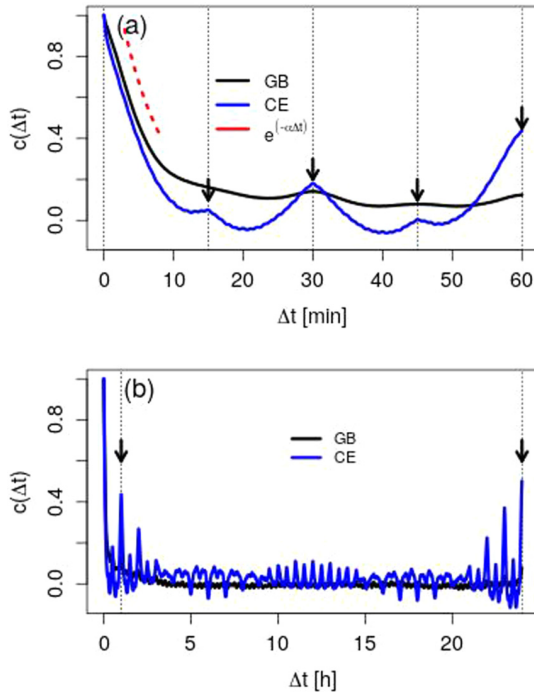


FIG. 3. Regular peaks in the autocorrelation demonstrate a mismatch between power supply and demand. (a) The autocorrelation $c(\Delta t)$ of GB and CE for a 1 h lag period. The black arrows indicate the times of trading/dispatch actions after 15 and 30 min, which cause the peaks in the autocorrelation. The dotted red line reports the exponential decay of the autocorrelation in the first 10 min. The inverse damping constants α^{-1} are estimated to be ~ 385 and ~ 312 s for the GB and CE power grids, respectively. (b) The autocorrelation function $c(\Delta t)$ of the GB (black) and CE (blue) data sets for a 24 h lag period. Regardless of regions, the initial exponential decay is followed by regular autocorrelation peaks. The black arrows highlight peaks of the autocorrelation after 1 h and also after 24 h, related to the periodicity of the frequency trajectory.

balance rapidly switches from positive to negative or vice versa, leading to large deviations of the grid frequency, which become visible in the autocorrelation function, see also [14]. In addition, daily routine, scheduled events, etc., contribute to an increased correlation every hour and 24 h, see black arrows in Fig. 3(b). Again, based on the specific regulations of different synchronous regions and their transmission system operators, the nature of the autocorrelation differs from region to region. For instance, the height of peaks in the GB autocorrelation in Fig. 3(a) is visibly smaller than CE, which we attribute to a smaller trading and regulatory volume and overall larger stochastic fluctuation in GB. Consequently, the deterministic aspect of the frequency dynamics is diluted in GB.

Finally, to clearly demonstrate the impact of the energy trading market and related regulator actions on the frequency, we show the daily average frequency of both GB and CE in Fig. 4. The daily average frequency for every second is obtained by averaging over all days of the year. The impact of the trading and regulation becomes clear, as we observe sharp frequency jumps upwards or downwards every hour in both GB and CE. The direction of the jump and thereby the

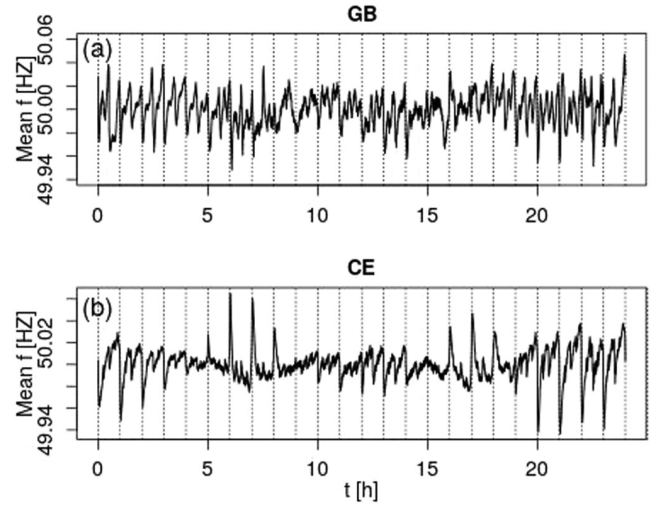


FIG. 4. Regular market activities induce periodic frequency jumps. Displayed is the frequency trajectory for (a) the GB grid and (b) CE grid, averaged over all 366 days in 2016. We notice clear frequency jumps every hour, consistent with the previous observation of peaks in the autocorrelation function.

question whether the grid is displaying a shortage or a surplus of power is not random but also follows a deterministic pattern. The market design is different for various synchronous grids or different countries within the same grid. For example, both the CE and the GB data display a periodicity of frequency jumps but the frequency dynamics within the CE grid appears more predictable. Frequency drops occur in the CE grid in each hour between 20:00 and 00:00, while the frequency clearly increases between 06:00 to 08:00 and 16:00 to 18:00. This pattern is linked to the slope of the demand curve. The steplike generation curve anticipates an increase or decrease of the demand [13]. In case of rising demand, such as during the morning, an increasing amount of power is dispatched for each trading interval, see Fig. 5(b). Every 15 min the generation is increased to anticipate the demand by the consumers. These discrete changes in the supplied power form the basis for the power mismatch in the synthetic frequency model discussed below.

III. STOCHASTIC PROPERTIES

Before we introduce a stochastic model for the power frequency dynamics, we perform some complementary tests to further characterize the underlying stochastic dynamics. Is the observed stochastic process stationary or nonstationary? Do we observe time symmetry, i.e., is the underlying process linear or nonlinear? Does the process depend on its past or only on the current state, i.e., is the process Markovian?

Stationary process. To test the reproducibility of the measured frequency, we first investigate the stationarity for the data. In the general definition, a probabilistic process is stationary if the probability of measured variables, in our case the probability of the frequency, does not depend on the time [38]. One of the standard methods to test the stationarity of a data set is analyzing its spectrum. The sharp peaks in Fig. 6 emphasise the existence of the periodicity on different

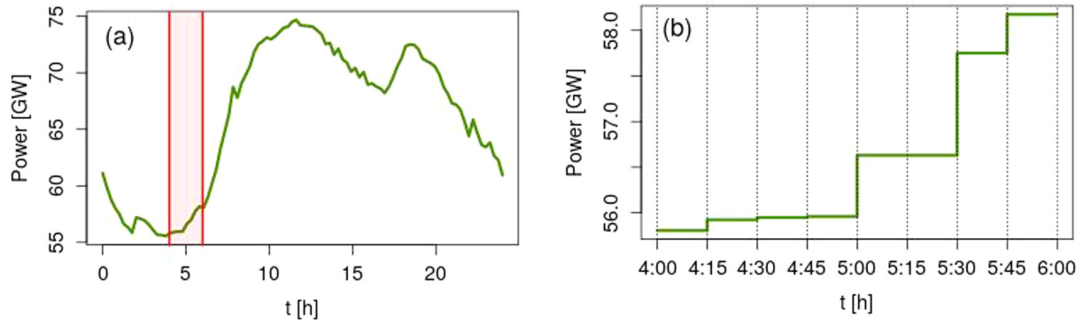


FIG. 5. Discrete power dispatch leads to jumps of the scheduled power supply. (a) We display the real dispatch trajectory of the electricity supply in Germany in one day in 2017 [37]. (b) The scheduled power jumps every 15 min, as highlighted by the zoom on the two hours highlighted in red in (a). Overall, the scheduled power supply approximates the changing demand throughout the day. Its discrete nature leads to jumps of the supply that has to be compensated by control mechanisms.

timescales in the considered data. According to the spectrum, there are visible periods every $1/4$, $1/2$, 1 , 12 , and 24 h in the grid frequency. This shows the nonstationary of the data on these timescales. However beyond 24 h, i.e., on longer timescales, the spectrum is decreasing and consequently the data becomes stationary.

There are other natural cycles influencing a power-grid system, such as the weekend-weekday pattern, as well as seasonal and yearly cycles. However, these cycles do not

seem to leave a significant imprint in the spectrum of the power-grid frequency. Our stochastic model will focus on the intermediate timescale and hence include the characteristic daily dispatch and demand pattern, while neglecting longer-term processes.

Linear process. Next, we investigate if there is any nonlinearity in the recorded power-grid frequency. For this purpose, consider the three-point autocorrelation of the frequency data as a measure of the time asymmetry in the data. If a time series is asymmetric in time, it is also nonlinear [38]. The following relations have been suggested to calculate the three-point autocorrelation for a data set [38]:

$$LT1 = \langle f(t)^2 f(t + \tau) \rangle - \langle f(t) f(t + \tau)^2 \rangle, \quad (3)$$

$$LT2 = \langle [f(t) - f(t + \tau)]^3 \rangle / \langle [f(t) - f(t + \tau)]^2 \rangle, \quad (4)$$

where LT stands for linear test. A linear, and therefore time-symmetric, trajectory has both $LT1$ and $LT2$ sufficiently close to zero. Checking the validity of our results for a realistic process, we compare the original data to a surrogate time series, that provides a reference point of $LT1$ and $LT2$ for a linear process. To generate the surrogate time series, we first take the Fourier transform (FT) of the original data and then randomize the phases. Finally, we employ an inverse FT to obtain the surrogate data. With the described procedure we suppress any nonlinearity in the process, and therefore the surrogate data includes only the linear characteristics of the considered data [39]. The original data is linear if the LT result of the original data lies within the value range of the LT results of the ensemble of surrogate data. Here, instead of displaying the full ensemble of surrogate data in Fig. 7, we have shown just an example for a surrogate data to avoid to obscure the figure. Comparing the $LT1$ results of the surrogate data sets with the $LT1$ of the original data sets displays that the qualitative behavior of both are equivalent, entailing that the processes approximately follow linear characteristics, for both the GB and the CE data sets, as seen in Fig. 7(a). Looking more closely at the $LT1$ for the CE surrogate data, which only includes the linear characteristics and fluctuations, we note that its deviation from zero are larger than $LT1$ for the original CE data. Investigating the value of $LT2$ for GB also confirms the linear characteristic of the data set. As the $LT1$ and $LT2$

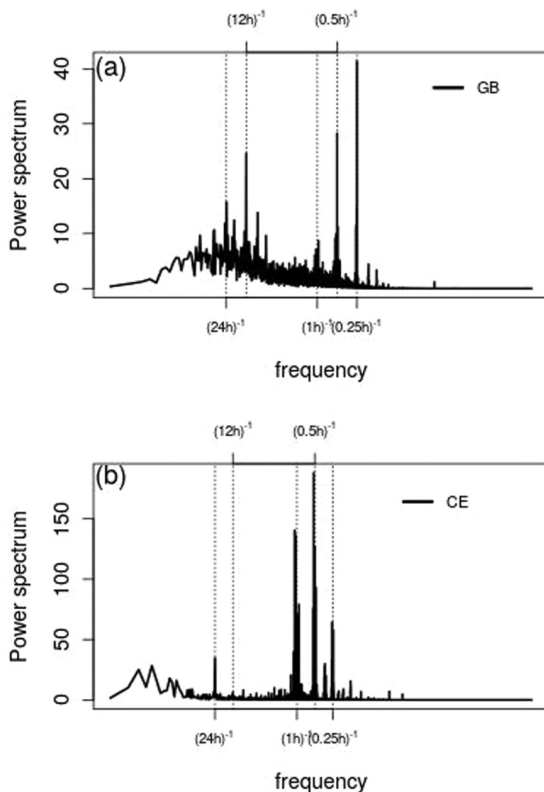


FIG. 6. Market activities and long timescales introduce nonstationarity. We plot the power spectrum of (a) GB and of (b) the CE data. The spectra exhibit well-determined peaks before they decay on a large timescale. The dotted vertical lines show $1/4$, $1/2$, 1 , 12 , and 24 h cycles (from right to left).

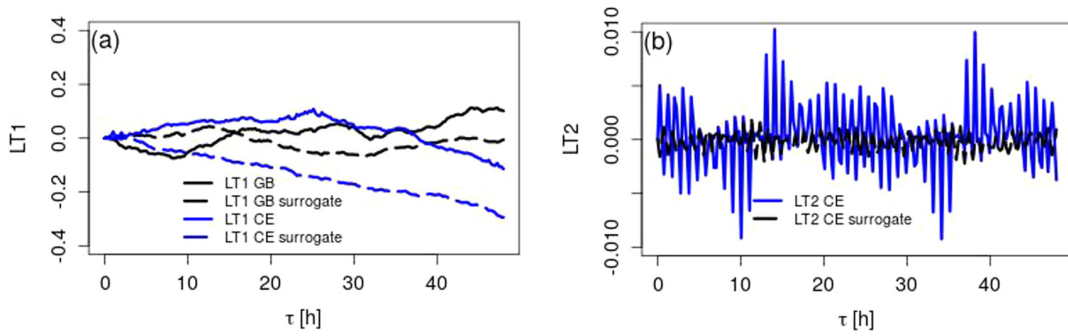


FIG. 7. The frequency trajectories display small nonlinear effects. (a) The $LT1$ results for the GB and CE frequency measurements. The dashed lines show the $LT1$ results for the surrogate data sets. The surrogate results act as a reference case of a linear model. Comparing the results of the original data with surrogate ones, we conclude both GB and CE are approximately linear. (b) $LT2$ results for the CE data. Surrogate (dashed black) and original data (solid blue) do differ more than when using $LT1$. This difference and the periodicity in $LT2$ are the signature of small nonlinear effects.

results for GB are the same, we only show the $LT1$ results. However, for the CE data set, $LT2$ indicates that the data might not be strictly linear but displays small nonlinearities, as seen in Fig. 7(b). As shown in Fig. 3, the effect of the market activity in CE is more regular and more severe than in GB, therefore we suspect that the nonlinearity in CE data is caused by the regular jumps in the frequency trajectory. When devising our model, we will therefore approximate the weakly nonlinear process as linear.

Chapman-Kolmogorov test. A fundamental property of stochastic processes is whether future states only depend on the current state or whether they have memory. In other words, whether the process is Markovian or not. A well-known approach to evaluate whether a process is Markovian is the Chapman-Kolmogorov test [36]. According to the Chapman-Kolmogorov test, the conditional PDFs of Markovian processes obey the following equation:

$$p(f_3, t_3 | f_1, t_1) = \int p(f_3, t_3 | f_2, t_2) p(f_2, t_2 | f_1, t_1) df_2, \quad (5)$$

where $t_3 > t_2 > t_1$. To test the Markovianity for the data, instead of employing directly Eq. (5), one considers its 2D and 3D conditional PDF. As shown in Fig. 8, $p(f_3, t_3 | f_1, t_1)$ and $p(f_3, t_3 | f_2, t_2; f_1, t_1)$ match approximately, implying the

power-grid frequency is mostly Markovian. Any stochastic model for the power frequency should therefore be Markovian as well.

IV. STOCHASTIC MODEL

We now introduce a synthetic model for the power-grid frequency as a stochastic, mostly linear, and Markovian process. The stochastic model presented here aims at reproducing essential features of a power grid, as well as its statistical characteristics, and consists of three independent systems: First, the intrinsic deterministic dynamics of the power grid, including primary and secondary control. Second, it embodies as well a stochastic signal or noise, as evidenced by the aforementioned frequency trajectories [27]. Third, we model the sudden power imbalance arising after the dispatch actions by implementing an appropriate deterministic function: We make use of historic dispatch data and apply it using a step function of the power. Other functions, such as artificial steps or sawtoothlike functions are also possible.

Instead of the actual frequency, we use the bulk angular velocity relative to the reference frequency of 50 Hz, $\omega = 2\pi(f - 50 \text{ Hz})$ to express our model. Contrary to network analysis on power grids [40,41], we have only access to

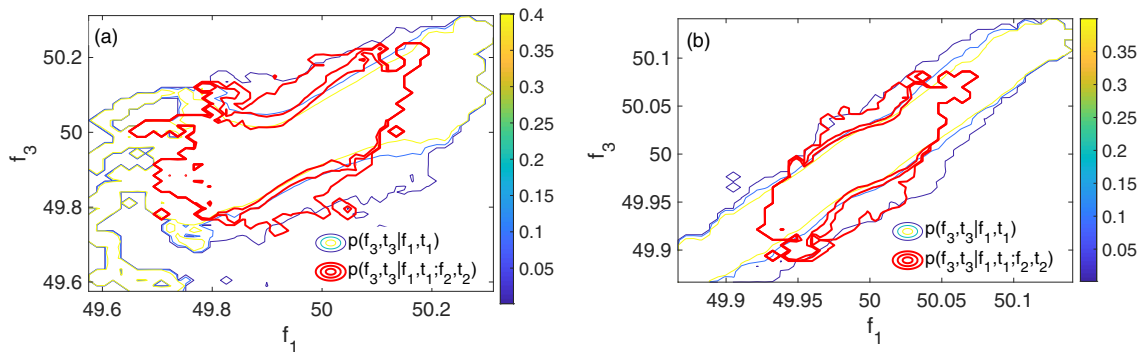


FIG. 8. The Markovian nature of the real data is confirmed by a Chapman-Kolmogorov test for (a) the GB grid and (b) the CE grid using the Baden-Württemberg data set. The proximity of the contour lines of $p(f_3, t_3 | f_1, t_1, f_2, t_2)$ (red contour) and $p(f_3, t_3 | f_1, t_1)$ (colored contour) show the validity of Chapman-Kolmogorov test for the frequency data sets. The time t_1 is chosen to contain ten data points to show the contours clearly. Next, the times t_2 and t_3 are multiples of t_1 , chosen as $2t_1$ and $3t_1$, respectively.

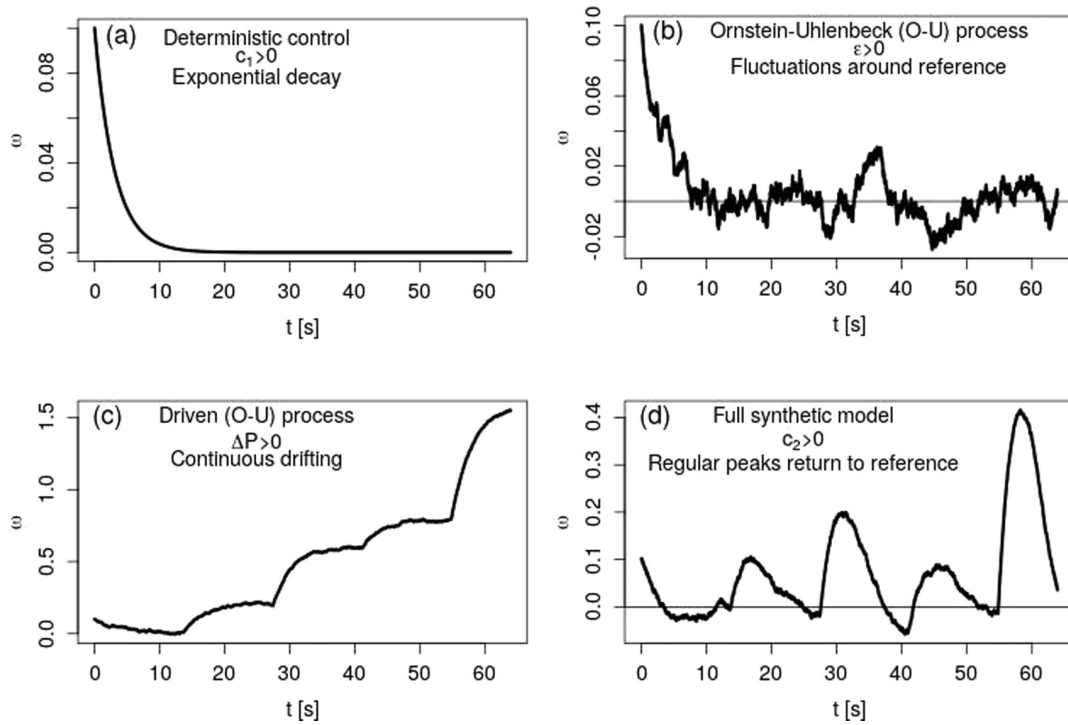


FIG. 9. All terms of the synthetic model (6) are necessary to reproduce the frequency trajectory. We plot the angular velocity ω as a function of time when using the synthetic frequency model (6) but setting individual parameters to 0. Parameters are chosen for pure illustrative purpose and we set $\omega(0) = 0.1$ as an initial condition. (a) Including only primary control leads to a pure exponential decay of the angular velocity. (b) Adding nonzero noise ϵ , we recover an Ornstein-Uhlenbeck process. (c) Including a step function for the power imbalance ΔP leads to a continuously drifting Ornstein-Uhlenbeck process. (d) Finally, including secondary control guarantees that the angular velocity returns back to the reference. Parameters are $\epsilon = 0.001/s^2$, $c_1 = 0.005/s$, $c_2 = 0.00003/s^2$, $\Delta P = 0.004/s^2$ at every hour and half or a quarter of it every 30 or 15 and 45 min, respectively.

frequency measurements on the global scale and therefore average over all nodes to obtain the averaged (bulk) frequency and angular velocity [42] $\omega = \frac{1}{M} \sum_{i=1}^N M_i \omega_i$, where $M = \sum_{i=1}^N M_i$ is the total inertia of all nodes and N is the number of nodes in the power grid. Typically, the frequency at each node is very close to the bulk frequency throughout the grid, with fluctuations indicating the gross power balance. Notable exceptions are high-frequency disturbances, which are typically localized [43,44], or interarea oscillations, where energy is oscillating from one part of the grid to another one. The synthetic model of the frequency dynamics is discussed in detail in [14]. It is given as a linear stochastic differential equation:

$$\frac{d\omega}{dt} = -c_1\omega - c_2\theta + \Delta P_{\text{ext}} + \epsilon\xi, \quad (6)$$

with bulk angle θ and its derivative $d\theta/dt = \omega$. Furthermore, ΔP_{ext} is the exogenous influence on the power balance, i.e., the trading or dispatch impact of the power imbalance, ϵ and ξ are the noise amplitude and Gaussian white noise function, respectively. Finally, c_1 is the magnitude of the fast-acting primary control, while c_2 is the magnitude of the secondary control which acts slower and lasts longer than primary control. We illustrate the contribution of the different terms of the synthetic model (6) in Fig. 9.

The full model is displayed in Fig. 9(d): In case of an abundance of generation, i.e., a sudden positive ΔP_{ext} , the

frequency increases above the reference (50 Hz). The primary control c_1 mitigates the sudden rise of the frequency and quickly stabilizes the frequency, but not at the nominal value of 50 Hz. Subsequently, the secondary control slowly restores the frequency back to its reference of 50 Hz. According to the time schedule of control systems, we assume that the primary control acts faster than secondary control, and consequently $c_1 \gg c_2$ [45,46].

Furthermore, the nature of the dispatch structure ΔP_{ext} must be specified. The generation of each power plant (the dispatch) is rapidly adapted by the operators, e.g., based on trading at the European Energy Exchange. As discussed in detail at the end of Sec. II, the operation of dispatchable power plants is scheduled at fixed intervals. As we have shown in Fig. 5 the power generation can increase or decrease every 15 min, which we model approximately as a step function, with potentially different step sizes at the 1 h, 30 min, or 15 min intervals. On the other hand, data of power generation in different regions or countries are generally available, and can be implemented directly in the model. In the model presented here, we extracted the power generation in Germany for the equivalent month of December 2017, and used this as the power balance ΔP_{ext} [37].

Before we compare results of the synthetic model with the real data, we need to determine suitable parameters. Details are given in [14] on how to estimate the parameters from a given frequency trajectory. In short, the noise amplitude

TABLE I. The parameters for the synthetic model for CE, December 2017.

ϵ (s^{-2})	c_1 (s^{-1})	c_2 (s^{-2})
0.00107	0.00915	0.00003

ϵ is estimated based on the stochastic fluctuations around the observed frequency trajectory, while the power imbalance ΔP_{ext} is directly read from the rise or sag of the frequency at the scheduled time points of dispatch, which are proportional to the missing or exceeding amount of power. (Notice that in our case we include the real power generation from Germany for December 2017, thus circumvent extracting the power generation ΔP_{ext} from the data.) Primary control c_1 is recovered by studying the process' affinity to revert its trajectory to the dispatched power and secondary control c_2 is estimated from the frequency recovery rate to the nominal value after a scheduled action [14].

V. QUANTITATIVE COMPARISON BETWEEN MODEL AND DATA

To evaluate the stochastic model described above, we generated one month of synthetic data with a 1-s resolution, mirroring the CE data from December 2017. The parameters for the synthetic model [14] are estimated from the 1-s resolution data series provided by [31] and their values are shown in Table I. The data for the power generation for the month of December 2017, in Germany, can be found in [37].

Now we repeat most of our statistical and stochastic analyses to compare how well the synthetic model reproduces the original data. First, we note that the general shape of the PDF [see Fig. 10(a)] and autocorrelation [see Fig. 10(b)] do agree well between the model (yellow) and the empirical data (black). Both the model and the data display heavy tails, i.e., the aforementioned deviation from Gaussianity. Furthermore, the autocorrelation function of the synthetic model captures the regular peaks, due to the changing dispatch. The decay of the autocorrelation function is approximately described by the current model. Both results emphasise the enormous impact of the energy market activity and dispatch structure on the dynamics and stability of the power system.

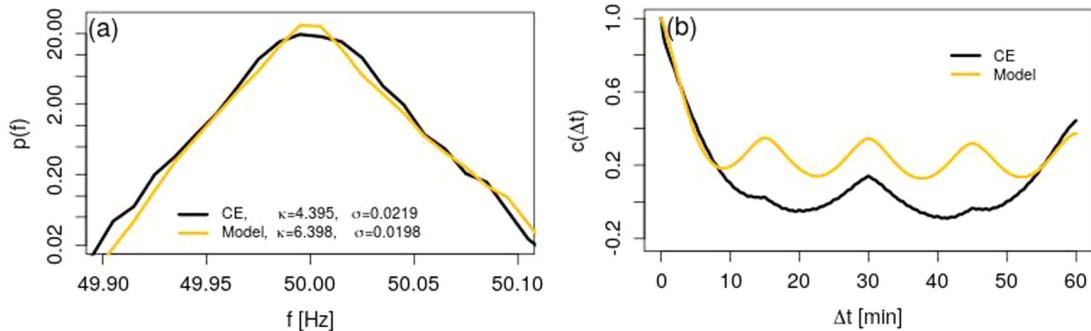


FIG. 10. The synthetic model captures important features of the real data, including trading peaks and heavy tails. (a) The probability distributions of the frequency data from CE in 2017 (black), compared to our synthetic model (yellow). Both display distinct heavy tails with kurtosis $\kappa > 3$. (b) The autocorrelation function of the frequency initially decays and then displays regular peaks at the trading intervals.

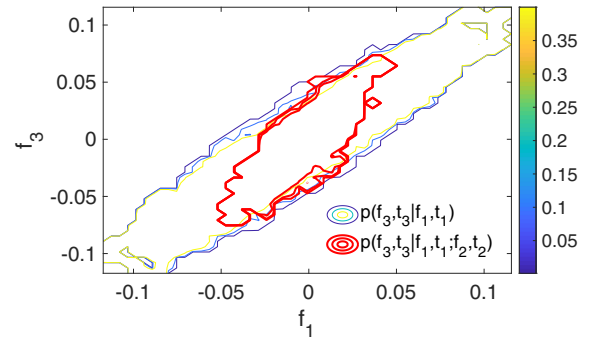


FIG. 11. Chapman-Kolmogorov test confirms the Markovian nature of the synthetic model. The test used 1 month of synthetic CE data generated by (6).

Consistent with our modeling assumptions, we find that the synthetic model is Markovian, based on a Chapman-Kolmogorov test, see Fig. 11. Similarly, we do observe that both *LT1* and *LT2* results show that the synthetic model has compatible characteristics with the real one, i.e., while the *LT1* reports a linear process, *LT2* results show a small nonlinearity in the synthetic, cf. Fig. 12. As we discussed in Sec. III, this nonlinear behavior is likely linked to the regular trading in the CE power grid.

We again emphasise that our model addresses the dynamics on the intermediate timescale of the frequency, i.e., approximately 30 s to a few hours. On shorter timescales, our model neglects: (i) dynamical behavior of rotating machines, (ii) nontrivial stochastic noise, (iii) network dynamics, and (iv) momentary reserve vs primary control. Moreover, the switching in trading is not instantaneous as we have assumed in the model. Similarly, our model does not include all effects acting on larger timescales, for example, (i) feed-in of wind and solar power, which determines how much inertia exists in the system and how much the generation side fluctuates, and (ii) dispatch of power plants determined on the spot market, such as the European Energy Exchange (EEX). This is especially relevant for areas where no historic market data are available or forecasts are attempted. In order to capture these effects, we would need a full fledged market model plus meteorological input for the weather data.

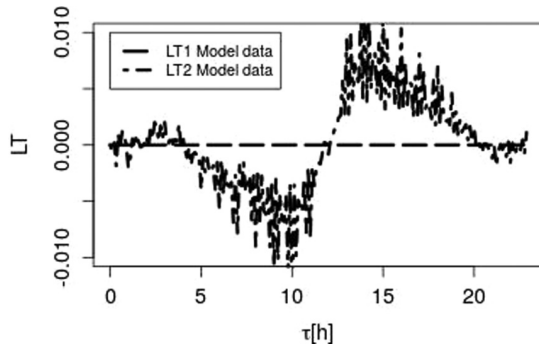


FIG. 12. The synthetic model is approximately linear. We apply the linear tests on the time series generated by the synthetic model: *LT1* shows linear characteristics for the CE data set, however *LT2* reports a small nonlinearity also found in the real data of the CE power-grid frequency.

The spectral analysis and the increment statistics of the synthetic data are shown in Fig. 13. Similarly to Fig. 2, in Fig. 13(a) the frequency increment statistics of the generated data also display non-Gaussian features on short timescales as the real data. The spectrum of the synthetic frequency trajectory displays several pronounced peaks, which are mostly

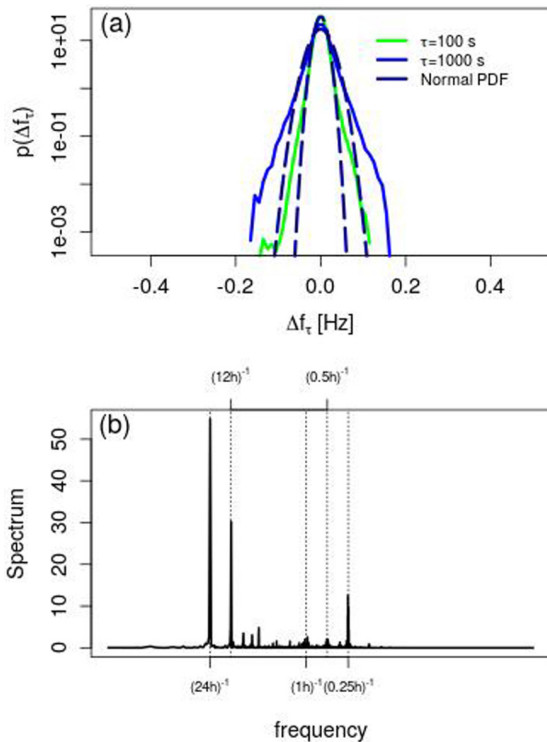


FIG. 13. Increment and spectral analysis of the synthetic model are consistent with the real data. (a) The increment statistics of the synthetic data shows non-Gaussian characteristics similar to the real one. (b) The spectrum of the synthetic frequency trajectory reports large peaks at the trading times, while it decays to zero for longer timescales. The dotted vertical lines show respectively, 1/4, 1/2, 1, 12, and 24 h from right to left.

consistent with the trading times of the model, i.e., 1/4, 12, and 24 h (cf. Fig. 13).

VI. DISCUSSION

In summary, we have presented an analysis of the statistics of power-grid frequency dynamics, with an emphasis on nonstandard behavior. In particular, we have shown the non-Gaussian nature of the power-grid frequency fluctuations in the aggregated and increments statistics, which includes heavy tails. Furthermore, we have demonstrated that the power-grid frequency trajectory is adequately described as a Markovian process and that it shows small nonlinear effects. Regulatory and trading events introduce some distinct periodicities in both autocorrelation and spectrum of the data sets. As we have mentioned before, the trading also has an obvious effect on the tails of frequency PDFs, or in other words, it is the source of non-Gaussianity in the measured data [27]. Finally, based on the observed properties, we have constructed a synthetic model that captures not only the aggregated statistics in terms of the histogram but also qualitatively reproduces the observed autocorrelation decay, correlation peaks due to market activity, increment statistics, and spectral properties of the real data [14]. The model is well suited to understand the energy-market effects on power-grid frequency on intermediate timescales and goes beyond previous studies focusing on a description [13,27] of trading or a stochastic theory [23]. We here focused on a statistical and stochastic analysis of real-world frequency dynamics, with a comparison to the presented model. The analysis of the synthetic model is consistent with our modeling assumptions, in that it is approximately Markovian and displays small nonlinear and periodic market effects. We should emphasise here that the observed heavy tails of the frequency distributions arise mainly due to trading actions, impacting not only the frequency temporally close to the market action but also several minutes later. This is clear since we only applied Gaussian noise to an otherwise linear dynamics. Only the deterministic trading actions can therefore cause the non-Gaussian properties. The spectral and increment properties of the synthetic model also approximate the original real-world data, which confirms again the effect of the trading market on the frequency dynamics. It is worth to reiterate that the presented model is conceptually simple, easy to implement, and includes a minimum set of adjustable parameters. Therefore, we explicitly did not model the machine dynamics, noise on very short timescales or a detailed market and dispatch model. Some alternative model approaches, involving more fitting parameters are explored in [14].

Concluding our analysis of power-grid frequency dynamics and the stochastic model we presented, including a structured comparison, may help to better understand the interplay of the internal dynamics and external disturbances of electric-power systems and to develop improved simulation models. A thorough understanding of this interplay is a prerequisite for the design and optimization of future electricity markets, as well as regulatory and control schemes. For instance, the current market design in the continental European grid regularly causes substantial frequency deviations when the dispatch is adjusted every 15 min such that primary control

has to be activated on a regular basis. A smoother change of the dispatch could reduce these frequency deviations and reduce stress onto the primary and secondary control system [13]. Alternatively, frequency regulations could be adapted in a way that the typical frequency deviations due to the changing dispatch are tolerated while exceptional cases are identified and handled by the control system. Our structured analyses (Markov, stationary, and linearity properties) and model may offer a powerful and versatile framework to study these questions, in particular because the model, while still simple, simultaneously captures essential features of the interplay of internal dynamics, control, and market activity. The presented analysis and modeling framework can thus contribute to the design of future power system, reducing the necessity for control actions and saving costs.

The model can further be used to assess the frequency stability of future power-grid structures, including in particular microgrids [8] or low-inertia grids [12]. Traditional dynamical stability analyses focus on local and global stability of fixed points and the impact of large isolated disturbances such as the sudden shutdown of the power plant. In comparison, the impact of ongoing stochastic disturbances on grid stability has received less attention. As evidenced in this study, the regulatory system and market design may have played an important role for these external stochastic effects.

We kept the model as simple as possible to reproduce key features of the frequency time series such as the histogram and the autocorrelation. Future research could naturally extend the model to better match the spectrum or long-time autocorrelation. Furthermore, one could investigate particular intervals of the power grid trajectory, e.g., high- vs low-demand intervals, such as weekdays vs weekends. Additional stochastic investigations could further quantify the agreement between real data and the synthetic model, e.g., by investigating higher-order N -point statistics, going beyond our current two-point statistics (increments).

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Federal Ministry of Education and Research (BMBF Grants No. 03SF0472F and No. 03EK3055F), the Helmholtz Association (via the joint initiative “Energy System 2050—A Contribution of the Research Field Energy” and Grant No. VH-NG-1025), and the German Science Foundation (DFG) by a grant toward the Cluster of Excellence “Center for Advancing Electronics Dresden” (cfaed). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 840825.

- [1] B. H. Obama, Presidential policy directive 21: Critical infrastructure security and resilience, Washington, DC (2013).
- [2] J. Markard, The next phase of the energy transition and its implications for research and policy, *Nat. Energy* **3**, 628 (2018).
- [3] M. Timme, L. Kocarev, and D. Witthaut, Focus on networks, energy and the economy, *New J. Phys.* **17**, 110201 (2015).
- [4] C. D. Brummitt, P. D. H. Hines, I. Dobson, C. Moore, and R. M. D’Souza, Transdisciplinary electric power grid science, *Proc. Natl. Acad. Sci. USA* **110**, 12159 (2013).
- [5] P. Kundur, *Power System Stability and Control* (McGraw-Hill, New York, 1994), Vol. 7.
- [6] ENTSO-E, Network code on requirements for grid connection applicable to all generators (rfg) (2013), <https://www.entsoe.eu/major-projects/network-code-development/requirements-for-generators/>.
- [7] J. Machowski, J. Bialek, and J. Bumby, *Power System Dynamics: Stability and Control* (John Wiley and Sons, Chichester, 2011).
- [8] X. Fang, S. Misra, G. Xue, and D. Yang, Smart Grids—The new and improved power grid: A survey, *Commun. Surveys Tutorials IEEE* **14**, 944 (2012).
- [9] P. Kotler, The Prosumer movement: A new challenge for marketers, *Adv. Consumer Res.* **13**, 510 (1986).
- [10] R. H. Lasseter and P. Paigi, Microgrid: A conceptual solution, in *2004 IEEE 35th Annual Power Electronics Specialists Conference (IEEE Cat. No. 04CH37551)*, Aachen, Germany (IEEE, 2004), Vol. 6, pp. 4285–4290.
- [11] P. C. Böttcher, A. Otto, S. Kettemann, and C. Agert, Time delay effects in the control of synchronous electricity grids, *Chaos* **30**, 013122 (2020).
- [12] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbić, Foundations and challenges of low-inertia systems, in *2018 Power Systems Computation Conference (PSCC)* (IEEE, New York, 2018).
- [13] T. Weißbach and E. Welfonder, High frequency deviations within the European power system—Origins and proposals for improvement, in *Proc. Power Syst. Conf. Expo., Seattle, WA, USA* (IEEE, 2009), pp. 1–6.
- [14] L. R. Gorjão *et al.*, Data-driven model of the power-grid frequency dynamics, in *IEEE Access* **8**, 43082 (2020), doi: [10.1109/ACCESS.2020.2967834](https://doi.org/10.1109/ACCESS.2020.2967834).
- [15] A. J. Wood, B. F. Wollenberg, and G. B. Sheblé, *Power Generation, Operation and Control* (John Wiley and Sons, New York, 2013).
- [16] A. Einfalt *et al.*, Energie der zukunft publizierbarer endbericht (2012), https://www.ea.tuwien.ac.at/fileadmin/t/ea/projekte/ADRES_Concept/PublizierbarerEndberichtADRES_815674.pdf.
- [17] T. Tjaden, J. Bergner, J. Weniger, and V. Quaschnig, Representative electrical load profiles of residential buildings in Germany with a temporal resolution of one second, ResearchGate: Berlin, Germany (2015), doi: [10.13140/RG.2.1.3713.1606](https://doi.org/10.13140/RG.2.1.3713.1606).
- [18] P. Milan, M. Wächter, and J. Peinke, Turbulent Character of Wind Energy, *Phys. Rev. Lett.* **110**, 138701 (2013).
- [19] M. Anvari *et al.*, Short term fluctuations of wind and solar power systems, *New J. Phys.* **18**, 063027 (2016).
- [20] C. Zhao and Y. Guan, Unified stochastic and robust unit commitment, *IEEE Trans. Power Syst.* **28**, 3353 (2013).
- [21] M. Anghel, K. A. Werley, and A. E. Motter, Stochastic model for power grid dynamics, in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on* (IEEE, New York, 2007), p. 113.
- [22] B. Schäfer, M. Matthiae, X. Zhang, M. Rohden, M. Timme, and D. Witthaut, Escape routes, weak links, and desynchronization

- in fluctuation-driven networks, *Phys. Rev. E* **95**, 060203(R) (2017).
- [23] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme, Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics, *Nat. Energy* **3**, 119 (2018).
- [24] H. Haehne, J. Schottler, M. Waechter, J. Peinke, and O. Kamps, The footprint of atmospheric turbulence in power grid frequency measurements, *Europhys. Lett.* **121**, 30001 (2018).
- [25] M. F. Wolff *et al.*, Heterogeneities in electricity grids strongly enhance non-Gaussian features of frequency fluctuations under stochastic power input, *Chaos* **29**, 103149 (2019).
- [26] K. Schmietendorf, O. Kamps, M. Wolff, P. G. Lind, P. Maass, and J. Peinke, Bridging between load-flow and Kuramoto-like power grid models: A flexible approach to integrating electrical storage units, *Chaos* **29**, 103151 (2019).
- [27] B. Schäfer, M. Timme, and D. Witthaut, Isolating the impact of trading on grid frequency fluctuations, in *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* (IEEE, New York, 2018), pp. 1–5.
- [28] Z. Li, O. Samuelsson, and R. Garcia-Valle, Frequency deviations and generation scheduling in the nordic system, in *PowerTech, 2011 IEEE Trondheim* (IEEE, New York, 2011), pp. 1–6.
- [29] National Grid, Frequency data (2014–2018), <http://www2.nationalgrid.com/Enhanced-Frequency-Response.aspx>.
- [30] Réseau de Transport d'Électricité (RTE), Network frequency (2014–2019), https://clients.rte-france.com/lang/an/visiteurs/vie/vie_frequence.jsp.
- [31] TransnetBW GmbH, Regelenergie Bedarf + Abruf (2019), <https://www.transnetbw.de/de/strommarkt/systemdienstleistungen/regelenergie-bedarf-und-abruf>.
- [32] B. Schäfer, D. Witthaut, M. Timme, and V. Latora, Dynamically induced cascading failures in supply networks, *Nat. Commun.* **9**, 1975 (2018).
- [33] K. Kashima, H. Aoyama, and Y. Ohta, Modeling and linearization of systems under heavy-tailed stochastic noise with application to renewable energy assessment, in *2015 54th IEEE Conference on Decision and Control (CDC)* (IEEE, New York, 2015), pp. 1852–1857.
- [34] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes. Stochastic Models with Infinite Variance* (Chapman and Hall, New York, 1994).
- [35] P. H. Westfall, Kurtosis as peakedness, 1905–2014 R.I.P., *Am. Stat.* **68**, 191 (2014).
- [36] C. Gardiner, *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences* (Springer, Berlin, 1985).
- [37] ENTSO-E, Generation Forecast—Day ahead, ENTSO-E <https://transparency.entsoe.eu/generation/r2/dayAheadAggregatedGeneration/show> (2019).
- [38] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997).
- [39] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Testing for nonlinearity in time series: The method of surrogate data, *Physica D* **58**, 77 (1992).
- [40] G. Filatrella, A. H. Nielsen, and N. F. Pedersen, Analysis of a power grid using a Kuramoto-like model, *Eur. Phys. J. B* **61**, 485 (2008).
- [41] M. Rohden, A. Sorge, M. Timme, and D. Witthaut, Self-Organized Synchronization in Decentralized Power Grids, *Phys. Rev. Lett.* **109**, 064101 (2012).
- [42] A. Ulbig, T. S. Borsche, and G. Andersson, Impact of low rotational inertia on power system stability and operation, *IFAC Proc.* **47**, 7290 (2014).
- [43] X. Zhang, S. Hallerberg, M. Matthiae, D. Witthaut, and M. Timme, Fluctuation-induced distributed resonances in oscillatory networks, *Sci. Adv.* **5**, eaav1027 (2019).
- [44] H. Haehne, K. Schmietendorf, S. Tamrakar, J. Peinke, and S. Kettemann, Propagation of wind-power-induced fluctuations in power grids, *Phys. Rev. E* **99**, 050301(R) (2019).
- [45] P. Kundur *et al.*, Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions, *IEEE Trans. Power Syst.* **19**, 1387 (2004).
- [46] E. B. T. Tchuisseu *et al.*, Curing Braess' paradox by secondary control in power grids, *New J. Phys.* **20**, 083005 (2018).

2.1.3 Publication #3

L. Rydin Gorjão and F. Meirinhos. `kramersmoyal`: Kramers–Moyal coefficients for stochastic processes. *Journal of Open Source Software* **4**(44), 2019, p. 1693, Ref. [3].

Status: published

kramersmoyal: Kramers–Moyal coefficients for stochastic processes

Leonardo Rydin Gorjão^{1, 2, 3, 4} and Francisco Meirinhos⁵

1 Department of Epileptology, University of Bonn, Venusberg Campus 1, 53127 Bonn, Germany **2** Helmholtz Institute for Radiation and Nuclear Physics, University of Bonn, Nußallee 14–16, 53115 Bonn, Germany **3** Forschungszentrum Jülich, Institute for Energy and Climate Research - Systems Analysis and Technology Evaluation (IEK-STE), 52428 Jülich, Germany **4** Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany **5** Physikalisches Institut and Bethe Center for Theoretical Physics, Universität Bonn, Nussallee 12, 53115 Bonn, Germany

DOI: [10.21105/joss.01693](https://doi.org/10.21105/joss.01693)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [Kristen Thyng](#) ↗

Reviewers:

- [@dawbarton](#)
- [@Shibabrat](#)

Submitted: 23 August 2019

Published: 19 December 2019

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

A general problem for evaluating Markovian stochastic processes is the retrieval of the moments or the Kramers–Moyal coefficients \mathcal{M} from data or time-series. The Kramers–Moyal coefficients are derived from an Taylor expansion of the master equation that describes the probability evolution of a Markovian stochastic process.

Given a set of stochastic data, ergodic or quasi-stationary, the extensive literature of stochastic processes awards a set of measures, such as the Kramers–Moyal coefficients or its moments, which link stochastic processes to a probabilistic description of the process or of the family of processes (Risken, 1996). Most commonly known is the Fokker–Planck equation or truncated forward Kolmogorov equation, partial differential equations, obtained from the Taylor expansion of the master equation.

Of particular relevance is the growing evidence that real-world data displays higher-order ($n > 2$) Kramers–Moyal coefficients, which has a two-fold consequence: The common truncation at third order of the forward Kolmogorov equation, giving rise to the Fokker–Planck equation, is no longer valid. The existence of higher-order ($n > 2$) Kramers–Moyal coefficients in recorded data thus invalidates the aforementioned common argument for truncation, thus rendering the Fokker–Planck description insufficient (Tabar, 2019). A clear and common example is the presence of discontinuous jumps in data (Aït-Sahalia, 2002; Anvari, Tabar, Peinke, & Lehnertz, 2016), which can give rise to higher-order Kramers–Moyal coefficients, as are evidenced in Gorjão, Heysel, Lehnertz, & Tabar (2019) and references within.

Calculating the moments or Kramers–Moyal coefficients strictly from data can be computationally heavy for long data series and is prone to inaccuracy especially where the density of data points is scarce, for example, usually at the boundaries on the domain of the process. The most straightforward approach is to perform a histogram-based estimation to evaluate the moments of the system at hand. This has two main drawbacks: it requires a discrete space of examination of the process and is shown to be less accurate than using kernel-based estimators (Lamouroux & Lehnertz, 2009).

This library is based on a kernel-based estimation, *i.e.*, the Nadaraya–Watson kernel estimator (Nadaraya, 1964; Watson, 1964), which allows for more robust results given both a wider range of possible kernel shapes to perform the calculation, as well as retrieving the results in a non-binned coordinate space, unlike histogram regressions (Silverman, 2018). It further employs a convolution of the time series with the selected kernel, circumventing the computational issue of sequential array summation, the most common bottleneck in integration time and computer memory.

The package presented here contains several options: A general open-source toolbox for the calculation of Kramers–Moyal coefficients for any given data series of any dimension and to any order, with a selection of commonly-used kernel estimators.

Mathematics

For a general N -dimensional Markovian process $\mathbf{x}(t) \in \mathbb{R}^N$ the Kramers–Moyal yields all orders of the cumulants of the conditional probability distribution $P(\mathbf{x}', t + \Delta T | \mathbf{x}, t)$ as

$$\mathcal{M}^\sigma(\mathbf{x}, t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int d\mathbf{x}' [\mathbf{x}(t)' - \mathbf{x}(t)]^\sigma P(\mathbf{x}', t + \Delta T | \mathbf{x}, t), \quad (1)$$

with $[\dots]^\sigma$ a dyadic multiplication and the power σ allowing for a set of powers depending on the dimensionality of the process (Risken, 1996).

The exact evaluation of the Kramers–Moyal coefficients for discrete or discretised datasets $\mathbf{y}(t)$ —any human measure of a process is discrete, as well as any computer generated data—is bounded by the timewise limit imposed. Taking as an example a two-dimensional case with $\mathbf{y}(t) = (y_1(t), y_2(t)) \in \mathbb{R}^2$, the Kramers–Moyal coefficients $\mathcal{M}^{[\ell, m]} \in \mathbb{R}^2$ take the form

$$\mathcal{M}^{[\ell, m]}(x_1, x_2, t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int dy_1 dy_2 (y_1(t + \Delta t) - y_1(t))^\ell (y_2(t + \Delta t) - y_2(t))^m P(y_1, y_2, t + \Delta t | x_1, x_2, t), \quad (2)$$

at a certain measure point (x_1, x_2) . The order of the Kramers–Moyal coefficients is given here by the superscripts ℓ and m .

Theoretically, there are still two details to attend to: Firstly, there is an explicit dependence on time t . For the case of stationary (or quasi-stationary) data discussed here, $P(\mathbf{x}', t + \Delta T | \mathbf{x}, t) = P(\mathbf{x}', \Delta T | \mathbf{x})$. This entails time-independent Kramers–Moyal coefficients $\mathcal{M}^\sigma(\mathbf{x})$. Secondly, Δt should take the limiting case of $\Delta t \rightarrow 0$ but the restriction of any measuring or storing device—or the nature of the observables themselves—permits only time-sampled or discrete recordings. In the limiting case where Δt is equivalent to the minimal sampling rate of the data, the Kramers–Moyal coefficients take the form, in our two-dimensional example, as

$$\mathcal{M}^{[\ell, m]}(x_1, x_2) = \frac{1}{\Delta t} \langle \Delta y_1^\ell \Delta y_2^m |_{y_1(t)=x_1, y_2(t)=x_2} \rangle, \text{ with } \Delta y_i = y_i(t + \Delta t) - y_i(t). \quad (3)$$

It is straightforward to generalise this to any number of dimensions. The relevance and importance of adequate time-sampling was extensively studied and discussed in Lehnertz, Zabawa, & Tabar (2018).

The Kramers–Moyal coefficients exist on an underlying probabilistic space, that is, there exists a probabilistic measure assigned to the process, stemming from the master equation describing the family of such processes. The conventional procedure, as mentioned previously, is to utilise a histogram regression of the observed process and retrieve, via approximation or fitting, the Kramers–Moyal coefficient. The choice of a histogram measure for the Kramers–Moyal coefficient results in an acceptable measure of the probability density functions of the process but requires a new mathematical space (a distribution space). The employment of a kernel-estimation approach, the Nadaraya–Watson estimator, implemented in this library, permits an identical overview without the necessity of a new (discretised) distribution space, given that the equivalent space of the observable can be taken.

Like the histogram approach for the measure of the Kramers–Moyal coefficients, each single measure of the observable $y(t)$ is averaged, with a designated weight, into the distribution space. The standing difference, in comparison to the histogram approach, is the removal of a (discrete) binning system. All points are averaged, in a weighted fashion, into the distribution space—aiding especially in cases where the number of point in a dataset is small—and awarding a continuous measurable space (easier for fitting, for example) (Lamouroux & Lehnertz, 2009).

Exemplary one-dimensional Ornstein–Uhlenbeck process

A one-dimensional Ornstein–Uhlenbeck process $y(t)$ takes the form

$$dy(t) = -\theta y(t)dt + \sigma dW(t), \quad (4)$$

with θ denoted as the *drift* or mean-reverting term, σ the *diffusion*, *volatility*, or stochastic amplitude, and $W(t)$ is a Brownian motion, *i.e.*, a Wiener process. For this particular example set $\theta = 0.3$ and $\sigma = 0.1$.

To be able to test the library and the retrieval on the Kramers–Moyal coefficients, and subsequently recover the drift and diffusion term, one can numerically integrate the process. We employ a Euler–Maruyama integrator, for simplicity. There are more reliable and faster integrators, for example JiTCSDE (Ansmann, 2018).

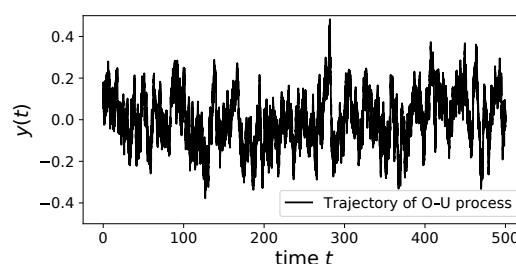


Fig. 1: Trajectory of Eq.(4) for $\theta = 0.3$ and $\sigma = 0.1$, for a total time of 500 time units, with a time step of 0.001, *i.e.*, comprising 5×10^5 data points.

For the present case, with an integration over 500 time units and with a timestep of 0.001, which can be seen in Fig. 1. The first and second Kramers–Moyal coefficients are presented in Fig. 2, where as well the conventional histogram-based estimation, a non-convolution based kernel estimation, and this library implementing a convolution of the kernel with the terms the right-hand side in Eq.(3). An Epanechnikov kernel was chosen for both kernel-based estimations.

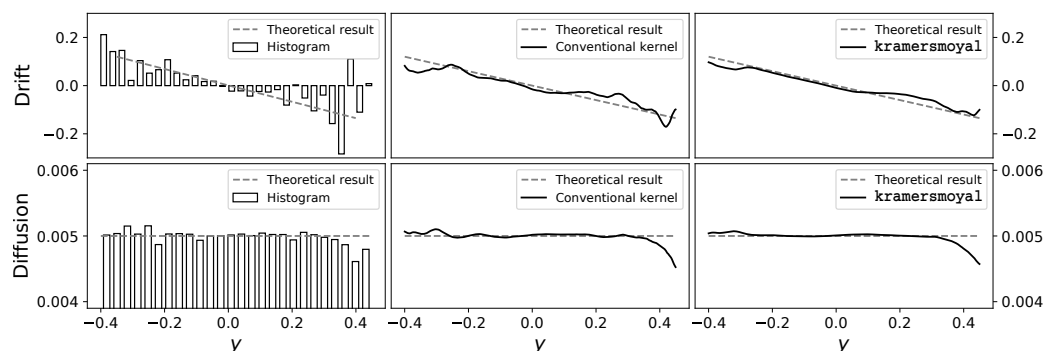


Fig. 2: Comparison of exemplary results of obtaining the Kramers–Moyal coefficients with a histogram-based approach, a conventional kernel-based approach, and the `kramersmoyal` library, sequentially left to right, from the numerical integration of Eq.(4). The top row displays the *drift* coefficient, *i.e.*, the first Kramers–Moyal coefficients. The bottom row

displays the *diffusion* coefficient, i.e., the second Kramers–Moyal coefficients. For the histogram 40 bins were used, for the conventional kernel and this library a space with 5500 numerical points were used, with a bandwidth of 0.05. The total number of points of the numerically integrated data is 5×10^5 .

Library

The presented `kramersmoyal` library is comprised of two separate blocks, `kernels` and `km`, and is a standalone package for a non-parametric retrieval of Kramers–Moyal coefficients, solely dependent on `numpy` and `scipy`. The sub-module `kernels` comprises the kernels for the kernel-based estimation, similarly available in `sklearn`, and `km` performs the desired Kramers–Moyal calculations to any desired power (Pedregosa et al., 2011). There exists a library to retrieve Kramers–Moyal coefficients in R (Rinn, Lind, Wächter, & Peinke, 2016).

In order to compare the computational speed up of the library the aforementioned Ornstein–Uhlenbeck Eq.(4) was used (with $\theta = 0.3$ and $\sigma = 0.1$), and the total time of integration of the process was increased iteratively. In Fig. 3 the comparative results of employing a histogram estimation with 200 bins, a conventional kernel-based regression in a space with 5500 numerical points, and this library’s kernel-convolution method, over similarly 5500 numerical points.

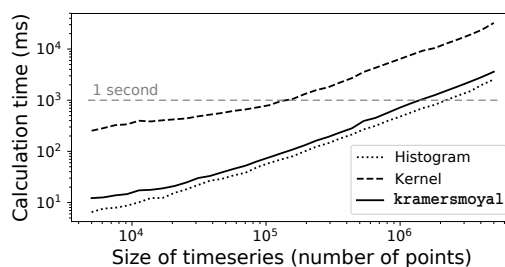


Fig. 3: Comparison of speed performance of obtaining the Kramers–Moyal coefficients with a histogram-based approach, a conventional kernel-based approach, and the `kramersmoyal` library, of a numerical integration of Eq.(4) over increasing number of data points. For the histogram 200 bins were used, for the conventional kernel and this library a space with 5500 numerical points was used. The total number of points of numerical integration was varied between 5×10^3 and 5×10^6 . The horizontal line indicates a total of 1 second. Integration performed on a laptop with an Intel Core i5 CPU @2.20~GHz (@2.56~GHz turbo).

Acknowledgements

L. R. G. and F. M. contributed equally to this project with their respective expertise. L. R. G. thanks Jan Heysel, Klaus Lehnertz, and M. Reza Rahimi Tabar for all the help in understanding stochastic processes and developing this package, Dirk Witthaut for the support during the process of writing and reviewing, Gerrit Ansmann for the help in understanding python’s intricacies, and Marieke Helmich for the text reviews. L. R. G. gratefully acknowledges support by the Helmholtz Association, via the joint initiative *Energy System 2050 - A Contribution of the Research Field Energy*, the grant No. VH-NG-1025, the scholarship funding from *E.ON Stipendienfonds*, and the *STORM - Stochastics for Time-Space Risk Models* project of the Research Council of Norway (RCN) No. 274410, under the supervision of Giulia di Nunno. F. M. gratefully acknowledges the fund, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 277625399 - CRC 185.

References

- Aït-Sahalia, Y. (2002). Telling from discrete data whether the underlying continuous-time model is a diffusion. *The Journal of Finance*, 57, 2075–2112. doi:[10.1111/1540-6261.00489](https://doi.org/10.1111/1540-6261.00489)
- Ansmann, G. (2018). Efficiently and easily integrating differential equations with JiTCODE, JiTCDDE, and JiTCSDE. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(4), 043116. doi:[10.1063/1.5019320](https://doi.org/10.1063/1.5019320)
- Anvari, M., Tabar, M. R. R., Peinke, J., & Lehnertz, K. (2016). Disentangling the stochastic behavior of complex time series. *Scientific reports*, 6, 35435. doi:[10.1038/srep35435](https://doi.org/10.1038/srep35435)
- Gorjão, L. R., Heysel, J., Lehnertz, K., & Tabar, M. R. R. (2019). Analysis and data-driven reconstruction of bivariate jump-diffusion processes. *Physical Review E*, (in press). Retrieved from <https://arxiv.org/abs/1907.05371>
- Lamouroux, D., & Lehnertz, K. (2009). Kernel-based regression of drift and diffusion coefficients of stochastic processes. *Physics Letters A*, 373(39), 3507–3512. doi:[10.1016/j.physleta.2009.07.073](https://doi.org/10.1016/j.physleta.2009.07.073)
- Lehnertz, K., Zabawa, L., & Tabar, M. R. R. (2018). Characterizing abrupt transitions in stochastic dynamics. *New Journal of Physics*, 20(11), 113043. doi:[10.1088/1367-2630/aaf0d7](https://doi.org/10.1088/1367-2630/aaf0d7)
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1), 141–142. doi:[10.1137/1109020](https://doi.org/10.1137/1109020)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from <http://www.jmlr.org/papers/v12/pedregosa11a.html>
- Rinn, P., Lind, P., Wächter, M., & Peinke, J. (2016). The Langevin approach: An R package for modeling Markov processes. *Journal of Open Research Software*, 4, e34. doi:[10.5334/jors.123](https://doi.org/10.5334/jors.123)
- Risken, H. (1996). *The Fokker–Planck Equation*. Springer, Berlin. doi:[10.1007/978-3-642-61544-3](https://doi.org/10.1007/978-3-642-61544-3)
- Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge, New York. doi:[10.1201/9781315140919](https://doi.org/10.1201/9781315140919)
- Tabar, M. R. R. (2019). *Analysis and data-based reconstruction of complex nonlinear dynamical systems*. Springer International Publishing. doi:[10.1007/978-3-030-18472-8](https://doi.org/10.1007/978-3-030-18472-8)
- Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4), 359–372. Retrieved from <http://www.jstor.org/stable/25049340>

2.2 Spatio-temporal analysis of power-grid frequency dynamics

2.2.1 Publication #4

L. Rydin Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer. *Open data base analysis of scaling and spatio-temporal properties of power grid frequencies*. Nature Communications **11**, p. 6362, 2020, Ref. [4].


Status: published

ARTICLE

<https://doi.org/10.1038/s41467-020-19732-7>

OPEN

Open database analysis of scaling and spatio-temporal properties of power grid frequencies

Leonardo Rydin Gorjão ^{1,2}, Richard Jumar ³, Heiko Maass³, Veit Hagenmeyer ³, G. Cigdem Yalcin ⁴, Johannes Kruse ^{1,2}, Marc Timme ⁵, Christian Beck ⁶, Dirk Witthaut ^{1,2} & Benjamin Schäfer ⁶✉

The electrical energy system has attracted much attention from an increasingly diverse research community. Many theoretical predictions have been made, from scaling laws of fluctuations to propagation velocities of disturbances. However, to validate any theory, empirical data from large-scale power systems are necessary but are rarely shared openly. Here, we analyse an open database of measurements of electric power grid frequencies across 17 locations in 12 synchronous areas on three continents. The power grid frequency is of particular interest, as it indicates the balance of supply and demand and carries information on deterministic, stochastic, and control influences. We perform a broad analysis of the recorded data, compare different synchronous areas and validate a previously conjectured scaling law. Furthermore, we show how fluctuations change from local independent oscillations to a homogeneous bulk behaviour. Overall, the presented open database and analyses constitute a step towards more shared, collaborative energy research.

¹Forschungszentrum Jülich, Institute for Energy and Climate Research-Systems Analysis and Technology Evaluation (IEK-STE), Jülich, Germany. ²Institute for Theoretical Physics, University of Cologne, Köln, Germany. ³Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics, Eggenstein-Leopoldshafen, Germany. ⁴Department of Physics, Istanbul University, 34134 Vezneciler, Istanbul, Turkey. ⁵Network Dynamics, Center for Advancing Electronics Dresden (cfaed) and Institute for Theoretical Physics, Technical University of Dresden, Dresden, Germany. ⁶School of Mathematical Sciences, Queen Mary University of London, London, UK. ✉email: b.schaefer@qmul.ac.uk

The energy system, and in particular the electricity system, is undergoing rapid changes due to the introduction of renewable energy sources, to mitigate climate change¹. To cope with these changes new policies and technologies are proposed^{2,3}, and a range of business models are implemented in various energy systems across the world⁴. New concepts, such as smart grids⁵, flexumers⁶, or prosumers⁷, are developed and tested in pilot regions. Still, studies rarely systematically compare different approaches, data, or regions, in part because freely available research data are lacking.

The frequency of the electricity grids is a key quantity to monitor, as it follows the dynamics of consumption and generation: a surplus of generation, e.g., due to an abundance of wind feed-in, directly translates into an increased frequency. Vice versa, a shortage of power, e.g., due to a sudden increase in demand, leads to a dropping frequency. Many control actions monitor and stabilise the power-grid frequency when necessary, so that it remains close to its reference value of 50 or 60 Hz⁸. Implementing renewable energy generators introduces additional fluctuations, as wind or photo-voltaic generation may vary rapidly on various timescales^{9–11} and reduces the overall inertia available in the grid¹². These fluctuations pose new research questions on how to design and stabilise fully renewable power systems in the future.

Analysis and modelling of the power-grid frequency and its statistics and complex dynamics have become increasingly popular in the interdisciplinary community, attracting much attention from mathematicians and physicists as well. Studies have investigated, e.g., different dynamical models^{13–15}, compared centralised vs. decentralised topologies^{16–18}, investigated the effect of fluctuations on the grid's stability^{19,20}, or how fluctuations propagate^{21,22}. Further research proposed real-time pricing schemes²³, optimised the placement of (virtual) inertia^{24,25}, or investigated cascading failures in power grids^{26–29}. However, these theoretical findings or predictions are rarely connected with real data of multiple existing power grids.

In addition to the need raised by theoretical models from the physics and mathematics community, there is also a great need for open databases and analyses from an engineering perspective. Although there exist databases of frequency time series, such as GridEye/FNET³⁰ or GridRadar (<https://gridradar.net/>), these databases are not open, which limits their value for the research community. In particular, different scientists with access to selected, individual types of data only, from grid frequencies to electricity prices, demand and consumption dynamics, cannot combine their data with these databases, thereby hindering to study more complex questions, such as the impact of price dynamics or demand control on system stability.

Hence, open empirical data are necessary to validate theoretical predictions, adjust models, and apply new data analysis methods. Furthermore, a direct comparison of different existing power grids would be very helpful when designing future systems that include high shares of wind energy, as they are already implemented in the Nordic grid, or by moving towards liberal markets, such as the one in Continental Europe. Proposals of creating small autonomous cells, i.e., dividing large synchronous areas into microgrids³¹ should be evaluated by comparing synchronous power grids of different size to estimate fluctuation and stability risks. In addition, cascading failures, spreading of perturbations, and other analyses of spatial properties of the power system may be evaluated by recording and analysing the frequency at multiple measurement sites.

In this study, we present an analysis of an open database for power-grid frequency measurements³² recorded with an Electrical Data Recorder (EDR) across multiple synchronous areas^{33,34}. Details on how the recordings were made are described in ref. ³², whereas we focus on an initial analysis and

interpretation of the recordings, which are publicly available (<https://osf.io/by5hu/>). First, we discuss the statistical properties of the various synchronous areas and observe a trend of decreasing fluctuation amplitudes for larger power systems. Next, we provide a detailed analysis of a synchronised wide-area measurement carried out in Continental Europe. We perform a detailed analysis showing that short time fluctuations are independent, whereas long timescale trends are highly correlated throughout the network. We extract the precise timescales on which the power-grid frequency transitions from localised to bulk dynamics. Finally, we extract inter-area oscillations emerging in the Continental European (CE) area. Overall, by establishing this database and performing a first analysis, we demonstrate the value of a data-driven analysis in an interdisciplinary context.

Results

Data overview. We recorded power-grid frequency time series using a Global Positioning System (GPS)-synchronised frequency acquisition device called EDR^{33,34}, providing similar data as a Phasor Measurement Unit would. Recordings were taken at local power sockets, which have been shown to give similar measurement results as that of monitoring the transmission grid with GPS time stamps³⁵ (see also ref. ³² for details on the data acquisition and a description of the open database). In addition, we received a 1-week measurement from the Hungarian TSO for the two cities Békéscsaba and Győr. We marked the locations of the measurement locations on a geographic map in Fig. 1a, b. Still, many more synchronous areas in the Americas, Asia, Africa, and Australia should be covered in the future.

To gain a first impression of the frequency dynamics, we visualise frequency trajectories in different synchronous areas and note quite a distinct behaviour (see Fig. 1c–e). We refer to each measurement by the country or state in which it was recorded (see also Supplementary Note 1). We group the measurements into (European) continental areas, (European) islands, and other (non-European) regions, which are also mostly continental. Most islands, such as Gran Canaria (ES-GC), Faroe Islands (FO), and Iceland (IS), but also South Africa (ZA), display large deviations from the reference frequency, whereas the continental areas, such as the Baltic (EE) and Continental European areas (DE), as well as the measurements taken in the United States (US-UT and US-TX) and Russia (RU), stay close to the reference frequency. There are still more differences within each group: e.g., the dynamics in ES-GC and ZA are much more regular than the very erratic jumps of the frequency over time observable in the FO and IS areas. Finally, we do not observe any qualitative difference between 50 and 60 Hz areas (right), when adjusting for the different reference frequency. It is noteworthy that some of the synchronous areas considered here are indeed coupled via high-voltage direct current (HVDC) lines but still possess independent synchronous behaviour. Specifically, the British (GB), Continental (DE), Baltic (EE), and Nordic (SE) European areas, as well as Mallorca (ES-PM), are connected in this way. The HVDC connection of Mallorca towards Continental Europe might be the reason it displays overall smaller deviations than the FO or IS areas, which cannot access another large synchronous area for balance.

Let us quantify the different statistics in a more systematic way by investigating distributions (histograms) and autocorrelation functions of the various areas. The distributions contain important information of how likely deviations from the reference frequency are, how large typical deviations are (width of the distribution), whether fluctuations are Gaussian (histogram displays an inverted parabola in log-scale), and whether they are skewed (asymmetric distribution). Analysing the distributions (histograms) of the individual synchronous areas (Fig. 2a–c), we

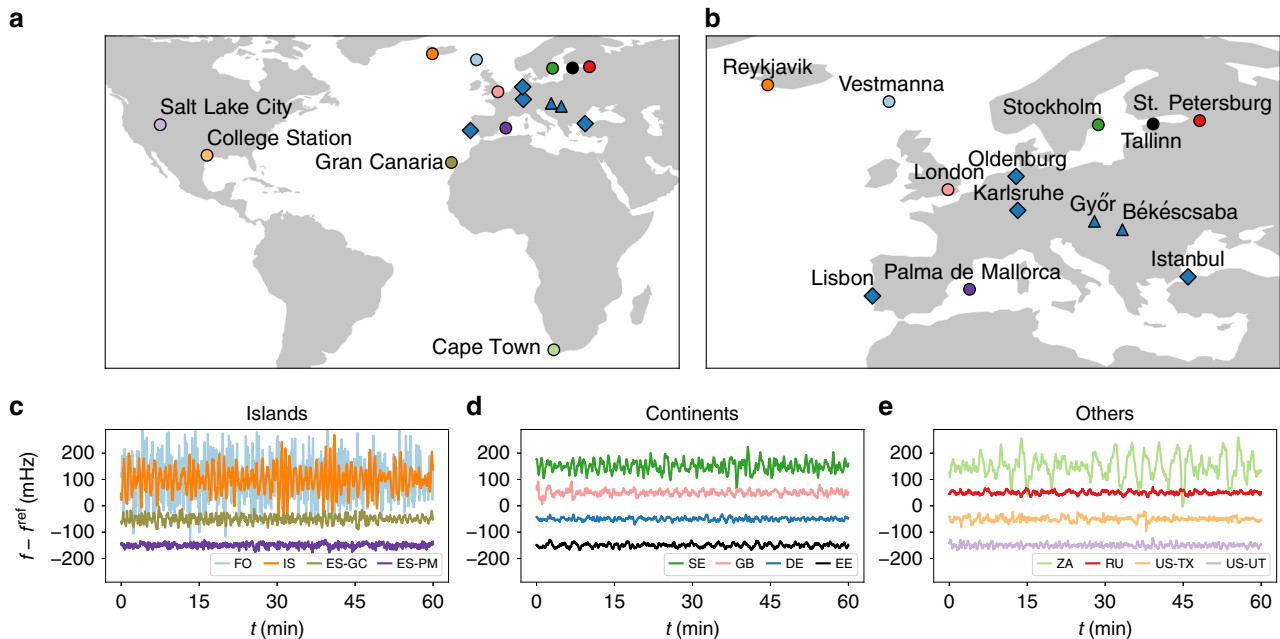


Fig. 1 Overview of available frequency data. **a** Different locations in Europe, Africa, and Northern America at which frequency measurements were taken. Australia and large parts of Asia are not displayed, as there were no measurements recorded. **b** Zoom of the European region (excluding Gran Canaria) with all locations labelled. Circles indicate measurement sites where single measurements for several days were taken, diamonds mark the four locations where we performed GPS-synchronised measurements, and triangles mark sites for which we received additional data. **c–e** Frequency trajectories display very different characteristics. We plot 1 h extracts of the deviations from the reference frequency of $f^{\text{ref}} = 50$ Hz (or 60 Hz for the US power grids), which are offset from the zero mean to improve readability. Panels **c–e** and following plots abbreviate the measurement sites using the ISO 3166 code for each country and each location is assigned a colour code, as in the maps in **a** and **b**. For more details on the data acquisition and measurement locations, see Supplementary Note 1 and ref. ³². Maps were created using Python 3 and geoplots.

note that the islands tend to exhibit broader and more heavy-tailed distributions than the larger continental areas. Still, there are considerable differences within each group. For example, we observe a larger standard deviation (SD) and thereby broader distribution in the Nordic (SE) and British (GB) areas compared to Continental Europe (DE), which is in agreement with earlier studies^{36,37}. Some distributions, such as those for Russia (RU) or the Baltic grid (EE), do show approximately Gaussian characteristics, whereas for several other areas, such as ES-GC and IS, they exhibit a high kurtosis ($\kappa^{\text{IS}} \approx 7$, as compared to $\kappa = 3$ for a Gaussian), i.e., heavy tails, and thereby a high probability for large frequency deviations. We provide a more detailed analysis of the first statistical moments, i.e., SD σ , skewness β , and kurtosis κ in Supplementary Note 1.

Complementary to the aggregated statistics observable in histograms, the autocorrelation function contains information on intrinsic timescales of the observed stochastic process (see Fig. 2d–f). For simple stochastic processes such as Ornstein–Uhlenbeck processes, we would expect an exponential decay $\exp(-\gamma\tau)$ of the autocorrelation with some damping constant γ ³⁸. Although most synchronous areas do show an approximately exponential decay, the decay constants vary widely. For example, the autocorrelation of the Icelandic data (IS) rapidly drops to zero, whereas the autocorrelation of the Nordic grid (SE) has an initial sharp drop, followed by a very slow decay. Other grids, such as the Faroe Islands (FO) or the Western Interconnection (US-UT) do show a slow decay, indicating long-lasting correlations, induced, e.g., via correlated noise. Finally, regular power dispatch actions every 15 min are clearly observable in the Continental European (DE), British (GB), and also the Mallorcan (ES-PM) grids, consistent with earlier findings^{36,37,39}.

In conclusion, we see that histograms are a good indicator of how heavy-tailed the frequency distributions are, whereas the

autocorrelation function reveals information on regular patterns and long-term correlations. These correlations are likely connected to market activity or regulatory action, demand and generation mixture, and other aspects specific to each synchronous area. Instead of going deep into individual comparisons, let us search for general applicable scaling laws instead.

Scaling of individual grids. For the first time, we have the opportunity to analyse numerous synchronous areas of different size, ranging from Continental Europe with a yearly power generation of about 3000 TWh⁴⁰ and a population of hundreds of millions to the Faroe Islands with a population of only tens of thousands. These various areas allow us to test a previously conjectured scaling law³⁶ of fluctuation amplitudes given as $\epsilon \sim 1/\sqrt{N}$, i.e., the aggregated noise amplitude ϵ in a synchronous area should decrease like the square root of the effective size of the area.

To derive this scaling relation, we formulate a stochastic differential equation of the aggregated frequency dynamics. A basic model, also known as the aggregated swing equation^{41,42}, is given as:

$$M \frac{d}{dt} \bar{\omega}(t) = -M\gamma \bar{\omega}(t) + \Delta P(t), \quad (1)$$

with bulk angular velocity $\bar{\omega}$, total inertia of a region M , power imbalance $\Delta P(t)$, and effective damping to inertia ratio γ , which also comprises primary control. The bulk angular velocity is the scaled deviation of the frequency from the reference: $\bar{\omega} = 2\pi(f - f^{\text{ref}})$ and $\Delta P(t)$ effectively represents noise acting on the system with mean $\langle \Delta P(t) \rangle = 0$, as generation and load are balanced on average. A simple scaling law for the frequency variability can be derived if the short-term power fluctuations at

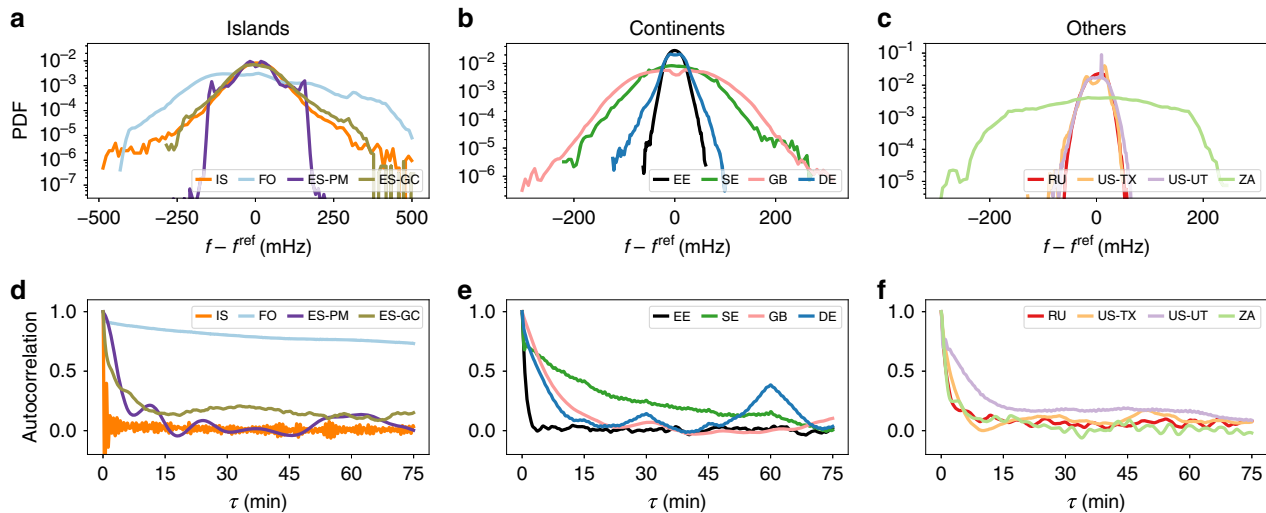


Fig. 2 Heterogeneity in power-grid statistics. Both histograms and autocorrelation functions display very distinct features between the different synchronous areas. **a–c** Histograms of the different synchronous areas provide insight on heavy tails but also the different scales of the fluctuations. We visualise the empirical probability distributions of the various areas by histograms on a logarithmic scale. **d–f** The complex autocorrelation decay reveals distinct timescales in the different grids. We compute the autocorrelation of each area for a time lag of up to 75 min.

each grid node are assumed to be Gaussian. If the grid has N nodes with identical noise amplitudes, the SD of the power imbalance scales as:

$$\sigma_{\Delta P} \sim \sqrt{N}. \quad (2)$$

At the same time, the total inertia typically scales linearly with the size of the grid, i.e., $M \sim N$. As a result, the amplitude of the total noise acting on the angular velocity dynamics scales as:

$$\epsilon \sim \frac{1}{M} \sigma_{\Delta P} \sim \frac{1}{\sqrt{N}}. \quad (3)$$

A more detailed derivation is provided in Supplementary Note 2 and discussed in refs. ^{36,37}. In addition, a technical discussion of extracting the aggregated noise amplitude is presented in ref. ⁴³. We note that the scaling law has to be modified if the noise at the nodes is not Gaussian³⁶.

To verify the proposed scaling law in Eq. (3), we approximate the number of nodes N by the population of an area, as generation data are not available for all synchronous areas, and population and generation tend to be approximately proportional⁴⁰. We utilise the population size as a proxy for size of the grid N . Indeed, we note that the aggregated noise amplitude ϵ does approximately decay with the inverse square root of the population size, as predicted (see Fig. 3). At a certain size, the noise saturates. The deviations from the prediction, such as by ZA and IS, are likely caused by different local control mechanisms, or non-Gaussian noise distributions, which we focus on in the next section. Interestingly, although FO and ES-PM do display non-Gaussian probability density functions, they follow the proposed scaling law. Why this is the case and how a fully non-Gaussian scaling law could capture this even better remain open questions for future work. Still, we observe a decay of the noise, approximately following the prediction over four orders of magnitude.

Increment analysis. In the previous section, we approximated the noise acting on each synchronous area as Gaussian to derive an approximate scaling law. In the following, we want to go beyond this simplification and investigate the rich short time statistics present in each synchronous area. We will see in particular how non-Gaussian distributions clearly emerge on the timescale of a few seconds.

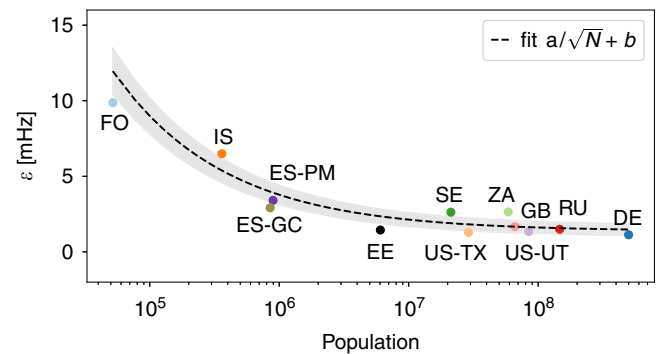


Fig. 3 The noise tends to decrease with an increasing size of the synchronous area until it saturates. We plot the extracted noise amplitude ϵ compared to the logarithm of the population in a given synchronous area. The population size serves as a proxy for the total generation and consumption of that area, as data on the size of the power grids is not commonly available. The shaded area is the SD of the ϵ estimation.

This short timescale is investigated via increments Δf_τ . The increment of a frequency time series is computed as the difference of two values of the frequency with a time lag τ :

$$\Delta f_\tau = f(t + \tau) - f(t). \quad (4)$$

An analysis of Δf_τ provides information on how the time series changes from one time lag τ to the next. On a short timescale of $\tau \approx 1$ s, the increments can be used as a proxy for the noise ϵ acting on the system (see also Supplementary Note 2).

The increments for a Wiener process, an often used reference stochastic process, are Gaussian regardless of the lag τ ³⁸. However, for many real-world time series, ranging from heart beats⁴⁴ and turbulence to solar and wind generation⁹, we observe non-Gaussian distributions for small lags τ . For many such processes with non-Gaussian increments, the probability distribution functions of the increments tend to approach Gaussian distributions for larger increments⁹. We observe a similar behaviour for the frequency statistics (see Fig. 4). The Nordic area (SE) displays deviations from Gaussianity for small lags τ but approximates a Gaussian distribution for larger τ . The Russian area (RU) even starts out with an almost Gaussian increment

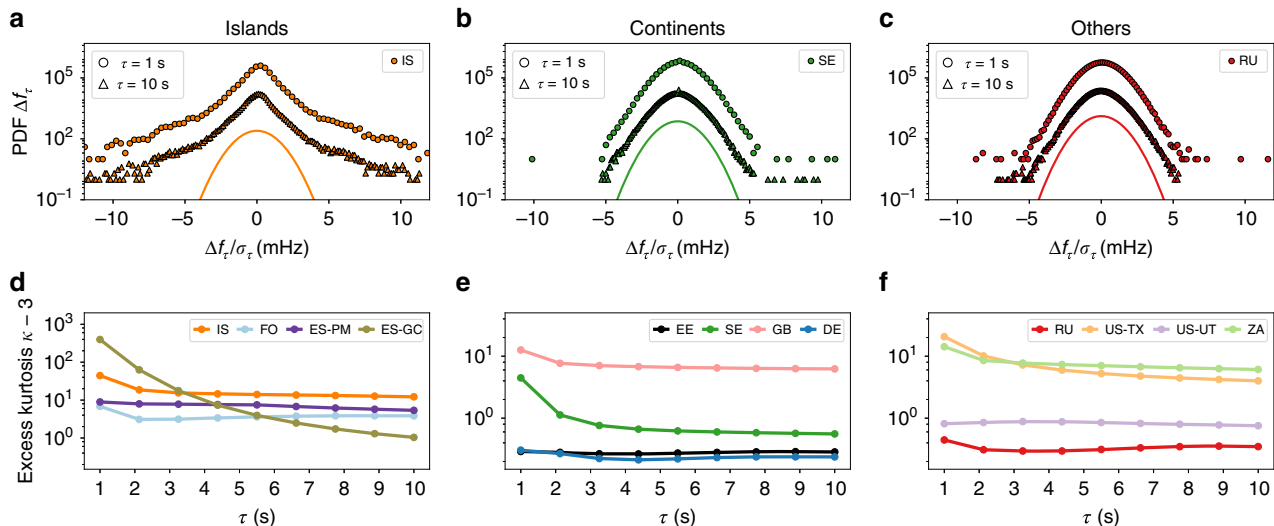


Fig. 4 Increment analysis reveals non-Gaussian characteristics dominantly in islands. **a–c** We display histograms of the increments Δf_τ for the lag values $\tau = 1, 10$ s for selected areas. The curves are shifted for visibility and compared to a Gaussian distribution as reference. **a** Iceland (IS) displays clear deviations from Gaussianity, even for larger increments τ . **b** The Nordic area (SE) displays a non-Gaussian distribution for $\tau = 1$ s, but approaches a Gaussian distribution for larger delays τ . **c** Russia (RU) has a Gaussian increment distribution for all lags τ . **d–f** We plot the excess kurtosis $\kappa - 3$ for the different examined power-grid frequency recordings on a log-scale. We observe a non-vanishing intermittency in Gran Canaria (ES-GC), Iceland (IS), Faroe Islands (FO), Mallorca (ES-PM), Britain (GB), Texas (US-TX), and South Africa (ZA). In contrast, the increments' distribution of the Baltic (EE), Continental Europe (DE), Nordic (SE), Russia (RU) synchronous areas, and the Western Interconnection (US-UT) approach a Gaussian distribution. See also Fig. 5 for an illustration how increments are computed from a trajectory.

distributions. Contrary, the Icelandic area (IS) shows clear deviations from a Gaussian distribution for all lags τ investigated here. Still, for larger lags, the pronounced tails flatten and the increment distribution slowly approaches a Gaussian distribution. The non-Gaussian increments on a short timescale point to non-Gaussian driving forces, e.g., in terms of generation or demand fluctuations acting on the power grid.

To investigate the deviations of the frequency increments from Gaussian properties, we utilise the excess kurtosis $\kappa - 3$ of the distribution. As the kurtosis κ , the normalised fourth moment of a distribution, is $\kappa = 3$ for a Gaussian distribution, a positive excess kurtosis points to heavy tails of the distribution.

Computing the excess kurtosis $\kappa - 3$ for all our data sets, we observe variable degrees of deviation across the various synchronous areas (Fig. 4). In some areas, the intermittent behaviour of the increments Δf_τ is subdued and the overall distribution approaches a Gaussian distribution (in EE, DE, SE, RU, and US-UT), i.e., the excess kurtosis $\kappa - 3$ becomes very small ($\lesssim 10^0$). In contrast, all islands as well as GB, US-TX, and ZA display large and non-vanishing intermittent behaviour, with a large excess kurtosis ($\sim 10^1 \dots 10^2$). IS and ES-GC show impressive deviations from Gaussianity, which require detailed modelling in the future.

We summarise that smaller regions tend to display more intermittency in their increments than larger regions, again consistent with findings on the scaling of the aggregated noise amplitude ϵ (Fig. 3). Furthermore, we observe that increment distributions tend to approach Gaussian distributions for larger increments, as expected⁹, but with distinct time horizons that depend on the grid area. For most of the islands the excess kurtosis remains high even for lags of 10 s. In contrast, in most areas of continental size, the excess kurtosis is very small already for lags larger than 1 s. Very interesting is also the following observation: non-Gaussian distributions in the aggregated frequency statistics (Fig. 2) are not necessarily linked with non-Gaussian increments. For example in Continental Europe (DE), we observe Gaussian increments but a non-Gaussian aggregated

distribution. The deviation from Gaussianity in the aggregated distribution, e.g., in terms of frequent extreme events, is likely explained by the external drivers, such as market activities⁴⁵. Finally, the analysis presented here extends previous increment analyses^{22,46}, which only considered increments of less than a second ($\tau < 1$ s), whereas we observe relevant non-Gaussian behaviour for larger increments ($\tau \geq 1$ s). We further analyse the differences between aggregated kurtosis and increment kurtosis in Supplementary Note 1, and discuss Castaing's model⁴⁷ and superstatistics⁴⁸ as more theoretical approaches towards increment analysis in Supplementary Note 3.

Correlated dynamics within one area. Moving away from comparing individual synchronous areas, we use GPS-synchronised measurements at multiple locations within the same synchronous area and the CE area, marked as diamonds and triangles, respectively, in Fig. 1. These measurements reveal that the frequency at different locations is almost identical on long timescales but differs on shorter timescales (see Fig. 5). Although the trajectories of the two German locations, Oldenburg and Karlsruhe, are almost identical, there are visible oscillations between the frequency values recorded in Central Europe (Karlsruhe) compared to the values recorded in the peripheries (Istanbul and Lisbon).

Let us quantify this by analysing the time series at the timescale of 1 s and hours (see Fig. 6). Increments Δf_τ , as also introduced above, reveal the short-term variability of a time series. In addition, we measure the long-term correlations on a timescale of hours by determining the rate of change of frequency (RoCoF). The RoCoF is the temporal derivative of the frequency and thereby very similar to increments. However, here it has a very different meaning, as we evaluate it only at every full hour and take into account several data points (see ref. ³⁷ and Methods). Thereby, the RoCoF mirrors the hourly power dispatch⁴⁹ and gives a good indication of long-term dynamics and deterministic external forcing. In the next section, we will also investigate the intermediate timescale of several seconds and inter-area oscillations.

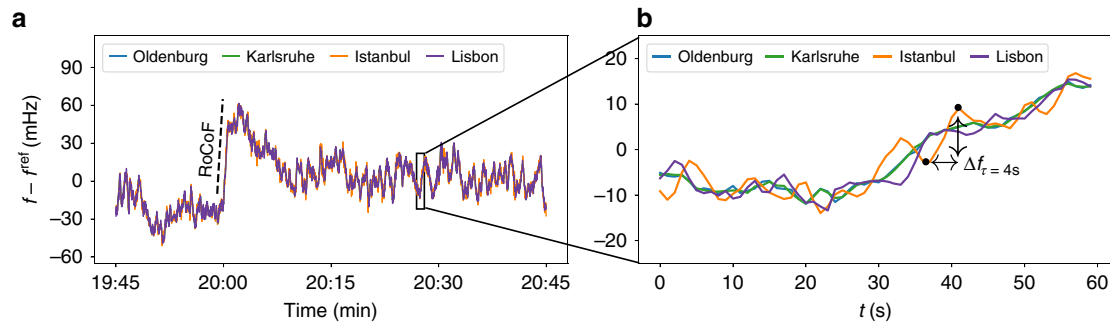


Fig. 5 Synchronised measurements within the Continental European (CE) synchronous area differ on the short timescale. We show a 1 h frequency trajectory recorded at four different sites in the CE area: Oldenburg, Karlsruhe, Lisbon, and Istanbul (**a**, **b**). We further illustrate the RoCoF (rate of change of frequency) as the slope of the frequency every hour and the increment statistics Δf_τ as the frequency difference between two points with time lag τ . For clarity, we do not include the two Hungarian measurement sites here, which produce similar results.

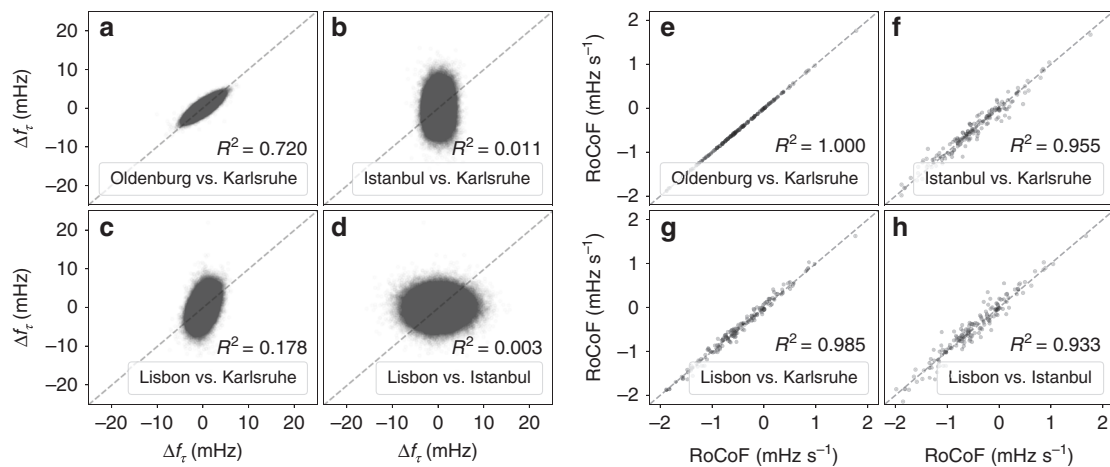


Fig. 6 From localised fluctuations to bulk behaviour. From left to right, we move our focus from short timescales (increments) to long timescales (RoCoF). **a–d** Short time increments are mostly independent. We compute the increment statistics $\Delta f_\tau = f(t + \tau) - f(t)$ for the increment time $\tau = 1$ s at the four sites in Continental Europe. The squared correlation coefficient R^2 is rounded to 4 digits. See also Supplementary Note 4 for larger lags τ and more details. **e–h** Correlations at the long timescale. We record the estimated rate of change of frequency (RoCoF) df/dt every 60 min for all four grid locations. In the scatter plots, each point represents the RoCoF or increment value Δf_τ computed at two different locations at the same time t .

Short timescale dynamics, as determined by frequency increments Δf_τ , are almost independent on the timescale of $\tau = 1$ s (see Fig. 6a–d). We generate scatter plots of the increment value $\Delta f_\tau(t)$ at the same time t at two different locations. If the increments are always identical, all points should lie on a straight line with slope 1. If the increments are completely uncorrelated, we would expect a circle or an ellipse aligned with one axis. Indeed, the increments taken at the same time for Oldenburg and Karlsruhe are highly correlated and almost always identical, i.e., the points in a scatter plots follow a narrow tilted ellipse (Fig. 6a). Moving geographically further away from Karlsruhe, the increments of Istanbul (Fig. 6b) are completely uncorrelated with those recorded in Karlsruhe, i.e., large frequency jumps in Istanbul may take place at the same time as small jumps happen in Karlsruhe. A similar picture of uncorrelated increments emerges when comparing Lisbon and Istanbul (Fig. 6d), whereas Lisbon vs. Karlsruhe displays some small correlation (Fig. 6c). At the two peripheral locations, Lisbon and Istanbul, the increment distributions are much wider, i.e., larger jumps on a short timescale are much more common in Istanbul and Lisbon than they are in Karlsruhe. For larger lags $\tau > 1$ s, the increments between all pairs become more correlated (see Supplementary Note 4).

Let us move to longer timescales. At the 60 min time stamps, power is dispatched in the CE grid to match the current demand, leading to a sudden surge in the frequency^{37,39,49}. Interestingly, the frequency dynamics at the different grid sites are very similar, i.e., the deterministic event of the power dispatch is seen

unambiguously everywhere in the synchronous area, almost regardless of distance (see Fig. 6e–h). All locations closely follow the same trajectory on the 1 h timescale. This is reflected in highly correlated RoCoF values, with a particularly good match between Oldenburg and Karlsruhe, and a linear regression coefficient of at least $R^2 \geq 0.93$ for all pairs (Fig. 6e–h).

We combine these different timescales in a single detrended fluctuation analysis (DFA), where we also integrate the two Hungarian locations (see Fig. 7). At short timescales, the DFA results differ for the six locations, while starting at the timescale of $t \sim 10^1$ s, the four curves coincide. For the timescale of 1 s, all locations are subject to different fluctuations, with Istanbul and Lisbon displaying the largest values of the fluctuation function. This is coherent with results of the increment analysis, where Istanbul and Lisbon have the broadest increment distributions (Fig. 6a–d). Moving to longer timescales of tens or hundreds of seconds, we observe a coincidence of the fluctuation function. This coincidence, i.e., identical behaviour for large timescales is in good agreement with the highly correlated RoCoF results (Fig. 6e–h). We may also interpret this change from short-term and localised dynamics to long-term and bulk behaviour as a change from stochastic to deterministic dynamics, i.e., the random fluctuations are localised and take place on a short timescale, whereas the deterministic dispatch actions and overall trends penetrate the whole grid on a long timescale. See also Methods and Supplementary Note 5 for details on the DFA methodology.

Spatio-temporal dynamics. Next, let us investigate the spatio-temporal aspect of the synchronised measurements. We connect the transition from local fluctuations towards bulk behaviour with the geographical distance of the measurement points, complementing earlier analysis based on voltage angles^{50,51}. We determine the typical time-to-bulk, i.e., the time necessary so that the dynamics at a given node approximates the bulk behaviour. To this end, we choose Karlsruhe, Germany, as our reference, which is very central within the CE synchronous area. The choice of the reference does not qualitatively change the results. For each of the remaining five locations, we compute the relative DFA function:

$$\eta(\ell) = \frac{F^2_{\text{Location}}(\ell) - F^2_{\text{Karlsruhe}}(\ell)}{F^2_{\text{Karlsruhe}}(\ell)} \quad (5)$$

with respect to Karlsruhe and ask, when does this difference drop below 0.1 (or 10%), i.e., when are the fluctuation at each location almost indistinguishable from the ones in Karlsruhe?

The further apart two locations are, the later they reach the bulk behaviour, i.e., the larger their time-to-bulk (see Fig. 8). This observation can be intuitively understood: two sites in close geographical vicinity are typically tightly coupled and can be synchronised by their neighbours, whereas sites far away have to stabilise on their own. Our time-to-bulk analysis quantifies this intuition. We consider both a linear and a quadratic fit. A linear dependence is expected if the bulk behaviour is realised by coupling via the shortest available path. In contrast, if the propagation is following a diffusive pattern via multiple

independent paths, we would expect a quadratic dependence of the time with respect to the distance. Indeed, the quadratic fit, following diffusive coupling, is a much better fit than a linear one, as indicated by a lower root-mean-squared-error 0.5, compared to 1.2 s in the linear case. Using the newly obtained fits, we find that a location only 100 km from Karlsruhe will have to independently stabilise fluctuations on the scale of 0.5–1 s and will then closely synchronise with the dynamics in Karlsruhe (our bulk reference). In contrast, a site 1000 km away has to stabilise already for about 3–5 s before it is fully integrated in the bulk. This gives additional guidance for the control within large synchronous areas, in particular for remote and weakly coupled sites. Clearly, these first estimates demonstrate that further research is necessary to validate and adjust spatio-temporal models of the power grid²¹.

Principal component analysis. So far, we have focused on when and how the localised fluctuations transition into a bulk behaviour. During this transition, on the intermediate timescale of about 5 s, we observe another phenomenon: ‘Inter-area oscillations’, i.e., oscillations between sites in different geographical areas far apart but still within one synchronous area. Different methods are available to extract spatial inter-area modes, ranging from Empirical Mode Decomposition⁵² to nonlinear Koopman modes⁵³. Here we use a principal component analysis (PCA)⁵⁴, which was already introduced to power systems when analysing inter-area modes and identifying coherent regions⁵⁵. A PCA separates the aggregated dynamics observed in the full system into ordered principal components, which we interpret as oscillation modes. Ideally, we can explain most of the observed dynamics of the full system by interpreting a few dominant modes. Each of these modes contains information of which geographical sites are involved in the modes dynamics, similar to an eigenvector. Typical behaviour includes a translational dynamics of all sites (the eigenvector with entries 1 everywhere) or distinct oscillations between individual sites (an eigenvector with entry 1 at one site and –1 at another site).

Indeed, applying a PCA to the synchronised measurements in CE, we can capture almost the entire dynamics with just three modes (see Fig. 9). In Fig. 9a, we provide the squared Fourier amplitudes of each mode and in Fig. 9b–d we visualise the first three modes geographically. These three modes already explain the largest shares λ_m of the total variance (see Supplementary Note 6 for the remaining modes and more details). The first mode (PC1) explains $\lambda_1 \approx 99.2\%$ of the variance and represents the synchronous bulk behaviour of the frequency. The second (PC2) and third (PC3) mode correspond to asynchronous inter-area modes. They contribute much less to the total variance due to their small amplitude (cf. Fig. 5). In PC2 (Fig. 9c), Western

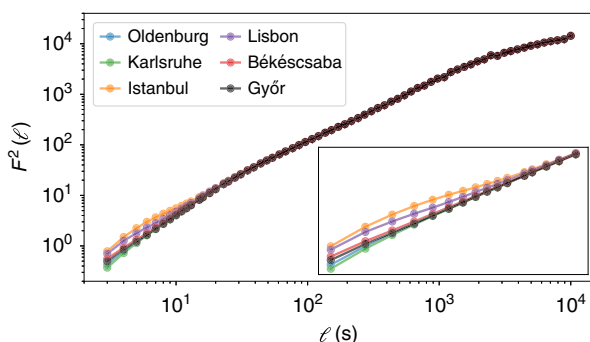


Fig. 7 Detrended Fluctuation Analysis (DFA) connects short and long timescales. We perform a DFA⁶⁵, with order $m=1$, in accordance with ref. ⁶⁴, and plot the fluctuation function $F^2(\ell)$ as a function of the time window length ℓ . The inset magnifies the values for $\ell \in \{10^0 \dots 2 \times 10^1\}$. The lines connect data points to each other to guide the eye.

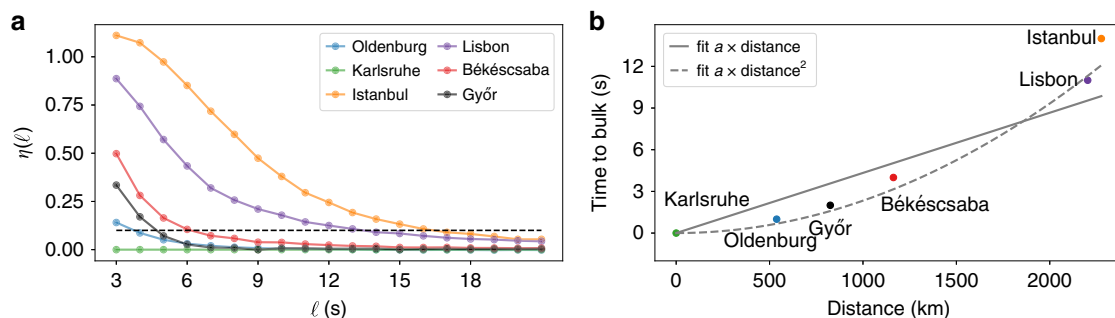


Fig. 8 The time-to-bulk increases with distance. **a** The relative DFA function $\eta(\ell)$ with Karlsruhe as reference; we determine the time-to-bulk as the time when this value reaches 0.1 (dashed line) (see also Eq. (5)). **b** We plot the so-extracted time-to-bulk vs. the distance from Karlsruhe and provide a simple fit for the first five points (i.e., excluding Istanbul, as it clearly behaves differently). We obtain a value for the linear fit $a = 5.2 \times 10^{-1} \text{ s km}^{-1}$, see Methods for details on the distance.

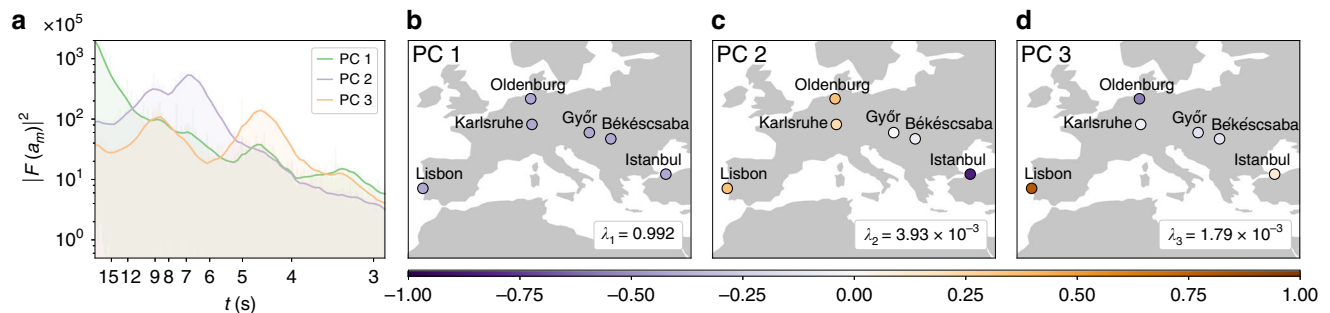


Fig. 9 Principal component analysis (PCA) of frequency recordings reveals inter-area oscillations. **a** In Continental Europe, the squared Fourier amplitudes $|F(a_m(t))|^2$ of the three dominant principal components (PCs) exhibits typical period lengths of $t \approx 7$ s and $t \approx 4.5$ s. **b** While the first spatial mode (PC1) corresponds to the bulk behaviour of the frequency and explains already $\lambda_1 = 99.2\%$ of the total variance, the second (PC2) and third (PC3) mode reveal asynchronous inter-area modes (**c**, **d**). We refer to Supplementary Note 6 for details on the method and the results.

Europe forms a coherent region that is in phase opposition to Istanbul (East–West dipole), whereas in PC3 (Fig. 9d), Lisbon and Istanbul swing in opposition to Oldenburg (North–South dipole). Similar results were found in an earlier theoretical study of the CE area, which also revealed global inter-area modes with dipole structures⁵⁶.

The temporal dynamics of the spatial modes exhibit typical frequencies of inter-area oscillations. Figure 9a shows the squared Fourier amplitudes $|F(a_m(t))|^2$ of the spatial modes. The components PC2 and PC3 have their largest peaks at $t \approx 7$ s and $t \approx 4.5$ s, which are the periods of these inter-area modes. These periods correspond well to the typical periods of inter-area oscillations, which are reported to be 1.25–8 s⁵⁷. On larger timescales $t > 12$ s, the amplitudes $|F(a_m(t))|^2$ of the inter-area modes drop below the values of PC1. Thus, the frequency dynamics is dominated by the bulk behaviour again, which is consistent with the estimated time-to-bulk of 12–15 s (Fig. 8).

Discussion

In this study, we have presented a detailed analysis of a recently published open database of power-grid frequency measurements³². We have compared various independent synchronous areas, from small regions, such as the FO and ES-PM areas, to large synchronous areas, such as the Western Interconnection in North America and the CE grid, spanning areas with only tens of thousand customers to those with hundreds of millions. Especially the smaller areas tend to show a larger volatility in terms of aggregated noise but also increment intermittency, such as IS and ES-GC. We have complemented this analysis of independent grids by GPS-synchronised measurements within the CE power grid, revealing high correlations of the frequency at long timescales but mostly independent dynamics on fluctuation-dominated short timescales. Compared to other studies applying synchronised, wide-area measurements, such as FNET/Grideye in the US³⁰ or evaluations from IS⁵⁸, the data we analysed here is freely available for further research³².

The comparison of different synchronous areas gives us a solid foundation to test previously conjectured scaling laws of fluctuations in power grids with their size³⁶, helps us to develop synthetic models³⁷, or predict the frequency⁵⁹ of small grids, such as microgrids. Furthermore, aggregating standardised measurements from different areas, we can compare countries with high shares of renewables (high hydro generation in Iceland or the Nordic area) with areas with almost no renewable generation (Mallorca) to learn how they influence the frequency dynamics and thereby the power-grid stability. Similarly, this comparison also gives insights on how different market structures impact the frequency statistics and stability of a power grid.

Our results on the spatial dependencies in the CE synchronous area are also highly relevant for the operation of power grids and other research in the field. The observations that the long-term behaviour is almost identical throughout the synchronous area but short time fluctuations differ, are in agreement with earlier theoretical findings²¹. Based on the DFA results (Figs. 7 and 8), we provide a quantitative estimate that at least for the CE area already at timescales of about 10 s, we observe an almost uniform bulk behaviour, even for locations thousands of kilometres apart. This bulk behaviour emerges much faster when locations are closer to one another.

In the regime of resonant behaviour²¹, we observe inter-area oscillations with period lengths of $t = 7$ s and $t = 4.5$ s, which we extract using a PCA. These timescales agree well with frequencies of inter-area oscillations reported in other studies in Europe^{56,57,60} but also in the United States⁶¹. However, we notice that the timescales separating bulk, resonance and local behaviour are different than the authors in a theoretical work²¹ assumed. There, local fluctuations were described for the 0.1 s timescale and bulk dynamics already started at times between about 2 and 5 s. This raises the question on how these timescales depend on the size and the dynamics of the power grid under consideration. Finally, we note that the PCA is a prime example for a model-free and data-driven analysis that leads to better understanding.

Our observation of frequency increments being independent on timescales of 1 s is consistent with earlier studies⁴⁶. For Continental Europe, we find that 1 s increments are correlated at small distances (below 500 km), but independent at locations far apart. On timescales of 1 s and below, we cannot observe global inter-area modes anymore. Instead, we expect local fluctuations that quickly decay with distance to their origin^{21,22}, which is consistent with our findings. The distribution of these short-term fluctuation was reported to exhibit a strongly non-Gaussian distribution when subject to intermittent wind power feed-in⁴⁶. In agreement with these results, the non-Gaussian effects vanish on timescales above 1 s in our recordings from Continental Europe. However, in other, particularly smaller, synchronous areas we even observe heavy-tailed increment distributions on timescales up to 10 s. This is likely related to the grid size and control regulations, although a detailed explanation still remains open.

In this study, we connect the mathematics and physics communities with the engineering community, by providing potent data analysis tools from the theoretical side and then connecting these findings in the practical domain of power-grid dynamics without the use of an explicit model. Both the data analysis and its interpretation could be very useful for the operation of individual grids. Our insights for the scaling could be used to improve control mechanisms, such as demand side management⁶², whereas our spreading insights give further indications about how

fast cascading failures will spread throughout the power grid²⁸. Several grid operators and other researchers have likely recorded power-grid frequency time series at many more grid locations than we could provide in this single study. All such recordings from different sources should be combined to enable more comparisons between the dynamics of synchronous areas of different sizes and under different conditions. The database studied here³² may offer a valuable starting point for such endeavours.

As data are still only scarcely available, there remain many open questions: can we systematically determine a propagation velocity of disturbances through the grid and compare these with theoretical predictions^{21,25,63}? Can we identify other time series influencing the power-grid frequency dynamics and quantify their correlation such as hydro power plants in the Nordic area or demand of aluminium plants in IC? Can we extract the impact of market activities on the frequency dynamics in all synchronous areas? From a statistical modelling perspective, it would be interesting to investigate the scaling of higher moments, i.e., skewness and kurtosis, with time lag and size in more detail. These questions constitute only a small selection from a multitude that an open database may help to address from a broad, interdisciplinary perspective, including engineering, mathematics, data science, time series analysis, and many other fields.

Methods

Data selection. We make use of the open database, described in detail in ref. ³², to perform all analyses presented in the main text and in Supplementary Notes 1–6. This data set contains recordings of 12 independent synchronous regions recorded between 2017 and 2020. Although some locations, such as the FO area only contain a single week of data, other regions, such as Continental Europe have been monitored for several months or years (for more details, see ref. ³²). However, due to some technical difficulties, e.g., loss of GPS signal or unplugging the device, some measurements are not a number, i.e., ‘NaN’, and are tagged as not reliable in the database. These entries have been deleted to compute the histograms and statistical measures in Supplementary Note 1. To compute the autocorrelation function and for the analysis of the synchronised measurement in Continental Europe, we selected the longest possible trajectory without any ‘NaN’ entries. As a final note, from the available ES-GC data, we are using the March 2018 data.

RoCoF computation. When determining the RoCoF, i.e., the time derivative of the frequency, we follow the same procedure as has been outlined in ref. ³⁷: we select a short time window centred around the anticipated dispatch jumps at 60 min of about 25 s length, i.e., starting at $(X) : 59 : 48$ and lasting until $(X + 1) : 00 : 12$ for all hours X . Then, we fit this short frequency trajectory with a linear function $f(t) = a + bt$. We are not interested in the offset a but the value of b gives us the slope of the frequency changes, i.e., the time derivative of the frequency is approximately given as $\frac{df}{dt} \approx b$.

Detrended fluctuation analysis. To carry out the DFA we follow a similar procedure as outlined in ref. ⁶⁴, using the package outlined in ref. ⁶⁵. The main idea is to detrend the data and extract the most dominant timescales by measuring the scaling behaviour of the data from increasing segments of data. The commonly denoted fluctuation function $F^2(\ell)$, function of the segment size ℓ on the time series, accounts for the variance of segmented data of increasing size. The scaling of the underlying process or processes can thus be extracted. In ref. ⁶⁴, a detailed study of the different timescales in power-grid frequencies can be found, largely focusing on scales of about 10 s and above, whereas we put particular emphasis on the smallest timescales available, of the order of 1 s. More details are given in Supplementary Note 5.

Time-to-bulk. To extract the time-to-bulk, seen in Fig. 8, we take the measurements of the DFA in Fig. 7 and utilise Karlsruhe as the reference for comparison. Having Karlsruhe as a reference, we compare the normalised fluctuations $\eta(\ell)$:

$$\eta(\ell) = \frac{F^2_{\text{location}}(\ell) - F^2_{\text{Karlsruhe}}(\ell)}{F^2_{\text{Karlsruhe}}(\ell)}, \quad (6)$$

(Eq. (5) in the main text), to extract the excess fluctuation at the different locations. As there is no standard, we choose a threshold value of 10% for fluctuations at the different recordings to be identical. Once $\eta(\ell)$ drops below this threshold of 10%, the data sets are considered to be identical. In this manner, we determine the time-to-bulk as the necessary time of a recording to exhibit the same fluctuation behaviour as the reference of Karlsruhe. The distance measures taken are the geographic distances with respect to Karlsruhe, applying OpenStreet Maps <https://www.openstreetmap.org/> and using the routing by Foot(OSRM). This yields the

following distances from Karlsruhe: Oldenburg: 538 km, Győr: 825 km, Békéscsaba: 1163 km, Lisbon: 2203 km, Istanbul: 2276 km. The reason to use route finding by foot is that the power grid is not taking any air plane routes but is limited also to the shortest routes available in the transmission grid. These distances in the power system might be even longer where transmission line density is low. It is noteworthy that our choice of geographical distance does not apply any assumption on the underlying power-grid topology. With full (yet currently unavailable) information about all operational transmission lines, a shortest path distance on the transmission network would be an alternative²².

Data availability

Frequency recordings are described in detail in ref. ³². An open repository containing all recordings can be accessed here: <https://osf.io/by5hu/>. The Hungarian TSO data are available here: <https://osf.io/m43tg/>. All data that support the results presented in the figures of this study are available from the authors upon reasonable request.

Code availability

Code to produce the presented analysis and figures is available on github: <https://github.com/LRydin>.

Received: 24 June 2020; Accepted: 22 October 2020;

Published online: 11 December 2020

References

- Murdock, H. E. et al. *Renewables 2020 Global Status Reports* (REN21, Paris, 2020).
- Meadowcroft, J. What about the politics? Sustainable development, transition management, and long term energy transitions. *Policy Sci.* **42**, 323 (2009).
- Markard, J. The next phase of the energy transition and its implications for research and policy. *Nat. Energy* **3**, 628–633 (2018).
- Rodríguez-Molina, J., Martínez-Núñez, M., Martínez, J.-F. & Pérez-Aguilar, W. Business models in the smart grid: challenges, opportunities and proposals for prosumer profitability. *Energies* **7**, 6142–6171 (2014).
- Fang, X., Misra, S., Xue, G. & Yang, D. Smart Grids - the new and improved power grid: a survey. *Commun. Surv. Tutor. IEEE* **14**, 944–980 (2012).
- Bärwaldt, G. Energy revolution needs interpreters. *ATZelektronik Worldw.* **13**, 68–68 (2018).
- Parag, Y. & Sovacool, B. K. Electricity market design for the prosumer era. *Nat. Energy* **1**, 16032 (2016).
- Kundur, P., Balu, N. J. & Lauby, M. G. *Power System Stability and Control* Vol. 7 (McGraw-Hill, New York, 1994).
- Anvari, M. et al. Short term fluctuations of wind and solar power systems. *New J. Phys.* **18**, 063027 (2016).
- Wolff, M. F. et al. Heterogeneities in electricity grids strongly enhance non-Gaussian features of frequency fluctuations under stochastic power input. *Chaos Interdiscip. J. Nonlinear Sci.* **29**, 103149 (2019).
- Wohland, J., Omrani, N. E., Keenlyside, N. & Witthaut, D. Significant multidecadal variability in German wind energy generation. *Wind Energy Sci.* **4**, 515–526 (2019).
- Hartmann, B., Vokony, I. & Táci, I. Effects of decreasing synchronous inertia on power system dynamics—overview of recent experiences and marketisation of services. *Int. Trans. Electr. Energy Syst.* **29**, e12128 (2019).
- Filatella, G., Nielsen, A. H. & Pedersen, N. F. Analysis of a power grid using a Kuramoto-like model. *Eur. Phys. J. B* **61**, 485–491 (2008).
- Schmietendorf, K., Peinke, J., Friedrich, R. & Kamps, O. Self-organized synchronization and voltage stability in networks of synchronous machines. *Eur. Phys. J. Spec. Top.* **223**, 2577–2592 (2014).
- Nishikawa, T. & Motter, A. E. Comparative analysis of existing models for power-grid synchronization. *New J. Phys.* **17**, 015012 (2015).
- Rohden, M., Sorge, A., Timme, M. & Witthaut, D. Self-organized synchronization in decentralized power grids. *Phys. Rev. Lett.* **109**, 064101 (2012).
- Menck, P. J., Heitzig, J., Marwan, N. & Kurths, J. How Basin stability complements the linear-stability paradigm. *Nat. Phys.* **9**, 89–92 (2013).
- Rodrigues, F. A., Peron, T. K. D., Ji, P. & Kurths, J. The Kuramoto model in complex networks. *Phys. Rep.* **610**, 1–98 (2016).
- Schäfer, B. et al. Escape routes, weak links, and desynchronization in fluctuation-driven networks. *Phys. Rev. E* **95**, 060203 (2017).
- Hindes, J., Jacquot, P. & Schwartz, I. B. Network desynchronization by non-Gaussian fluctuations. *Phys. Rev. E* **100**, 052314 (2019).
- Zhang, X., Hallerberg, S., Matthiae, M., Witthaut, D. & Timme, M. Fluctuation-induced distributed resonances in oscillatory networks. *Sci. Adv.* **5**, eaav1027 (2019).
- Hähne, H., Schmietendorf, K., Tamrakar, S., Peinke, J. & Kettemann, S. Propagation of wind-power-induced fluctuations in power grids. *Phys. Rev. E* **99**, 050301 (2019).

23. Schäfer, B., Matthiae, M., Timme, M. & Witthaut, D. Decentral smart grid control. *New J. Phys.* **17**, 015002 (2015).
24. Poolla, B. K., Bolognani, S. & Dörfler, F. Optimal placement of virtual inertia in power grids. *IEEE Trans. Autom. Control* **62**, 6209–6220 (2017).
25. Pagnier, L. & Jacquod, P. Inertia location and slow network modes determine disturbance propagation in large-scale power grids. *PLoS ONE* **14**, e0213550 (2019).
26. Simonsen, I., Buzna, L., Peters, K., Bornholdt, S. & Helbing, D. Transient dynamics increasing network vulnerability to cascading failures. *Phys. Rev. Lett.* **100**, 218701 (2008).
27. Yang, Y., Nishikawa, T. & Motter, A. E. Small vulnerable sets determine large network cascades in power grids. *Science* **358**, eaan3184 (2017).
28. Schäfer, B. & Yalcin, G. C. Dynamical modeling of cascading failures in the Turkish power grid. *Chaos Interdiscip. J. Nonlinear Sci.* **29**, 093134 (2019).
29. Nesti, T., Zocca, A. & Zwart, B. Emergent failures and cascades in power grids: a statistical physics perspective. *Phys. Rev. Lett.* **120**, 258301 (2018).
30. Chai, J. et al. Wide-area measurement data analytics using FNET/GridEye: a review. In *2016 Power Systems Computation Conference* (IEEE, Genoa, 2016).
31. Lasseter, R. H. & Paigi, P. Microgrid: a conceptual solution. In *2004 IEEE 35th Annual Power Electronics Specialists Conference* Vol. 6, 4285–4290 (IEEE, Aachen, 2004).
32. Jumar, R., Maass, H., Schäfer, B., Gorjão, L. R. & Hagenmeyer, V. Power grid frequency data base. Preprint at <https://arxiv.org/abs/2006.01771> (2020).
33. Maass, H. et al. First evaluation results using the new electrical data recorder for power grid analysis. *IEEE Trans. Instrum. Meas.* **62**, 2384–2390 (2013).
34. Maass, H. et al. Data processing of high-rate low-voltage distribution grid recordings for smart grid monitoring and analysis. *EURASIP J. Adv. Signal Process.* **2015**, 14 (2015).
35. Kakimoto, N., Sugumi, M., Makino, T. & Tomiyama, K. Monitoring of interarea oscillation mode by synchronized phasor measurement. *IEEE Trans. Power Syst.* **21**, 260–268 (2006).
36. Schäfer, B., Beck, C., Aihara, K., Witthaut, D. & Timme, M. Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics. *Nat. Energy* **3**, 119–126 (2018).
37. Rydin Gorjão, L. et al. Data-driven model of the power-grid frequency dynamics. *IEEE Access* **8**, 43082–43097 (2020).
38. Gardiner, C. W. *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences* (Springer, Germany, 1985).
39. Anvari, M. et al. Stochastic properties of the frequency dynamics in real and synthetic power grids. *Phys. Rev. Res.* **2**, 013339 (2020).
40. European Network of Transmission System Operators for Electricity. *Statistical Factsheet 2018* (ENTSO-E, Brussels, 2018).
41. Machowski, J., Bialek, J. & Bumby, J. *Power System Dynamics: Stability and Control* (Wiley, Chichester, 2011).
42. Ulbig, A., Borsche, T. S. & Andersson, G. Impact of low rotational inertia on power system stability and operation. *IFAC Proc. Volumes* **47**, 7290–7297 (2014).
43. Rydin Gorjão, L. & Meirinhos, F. kramersmoyal: Kramers–Moyal coefficients for stochastic processes. *J. Open Source Softw.* **4**, 1693 (2019).
44. Peng, C.-K. et al. Long-range anticorrelations and non-Gaussian behavior of the heartbeat. *Phys. Rev. Lett.* **70**, 1343 (1993).
45. Schäfer, B., Timme, M. & Witthaut, D. Isolating the impact of trading on grid frequency fluctuations. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* 1–5 (IEEE, Sarajevo, 2018).
46. Hähne, H., Schottler, J., Wächter, M., Peinke, J. & Kamps, O. The footprint of atmospheric turbulence in power grid frequency measurements. *Europhys. Lett.* **121**, 30001 (2018).
47. Castaing, B., Gagne, Y. & Hopfinger, E. Velocity probability density functions of high Reynolds number turbulence. *Phys. D Nonlinear Phenom.* **46**, 177–200 (1990).
48. Beck, C. & Cohen, E. G. D. Superstatistics. *Phys. A* **322**, 267–275 (2003).
49. Weißbach, T. & Welfonder, E. High frequency deviations within the European power system—origins and proposals for improvement. *VGB Powertech.* **89**, 26 (2009).
50. Cresap, R. & Hauer, J. Emergence of a new swing mode in the western power system. In *IEEE Transact. Power Apparatus and Systems* 2037–2045 (IEEE, 1981).
51. Chompoobutgool, Y. & Vanfretti, L. Identification of power system dominant inter-area oscillation paths. *IEEE Trans. Power Syst.* **28**, 2798–2807 (2012).
52. Messina, A. R. & Vittal, V. Extraction of dynamic patterns from wide-area measurements using empirical orthogonal functions. *IEEE Trans. Power Syst.* **22**, 682–692 (2007).
53. Susuki, Y. & Mezic, I. Nonlinear Koopman modes and coherency identification of coupled swing dynamics. *IEEE Trans. Power Syst.* **26**, 1894–1904 (2011).
54. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2007).
55. Anaparthi, K., Chaudhuri, B., Thornhill, N. & Pal, B. Coherency identification in power systems through principal component analysis. *IEEE Trans. Power Syst.* **20**, 1658–1660 (2005).
56. Grebe, E., Kabouris, J., Lopez Barba, S., Sattinger, W. & Winter, W. Low frequency oscillations in the interconnected system of Continental Europe. In *IEEE PES General Meeting* 1–7 (IEEE, Minneapolis, 2010).
57. Klein, M., Rogers, G. J. & Kundur, P. A fundamental study of inter-area oscillations in power systems. *IEEE Trans. Power Syst.* **6**, 914–921 (1991).
58. Tuttlberg, K., Kilter, J., Wilson, D. & Uhlen, K. Estimation of power system inertia from ambient wide area measurements. *IEEE Trans. Power Syst.* **33**, 7249–7257 (2018).
59. Kruse, J., Schäfer, B. & Witthaut, D. Predicting the power grid frequency. *IEEE Access* **8**, 149435–149446 (2020).
60. Vanfretti, L., Bengtsson, S., Perić, V. S. & Gjerde, J. O. Spectral estimation of low-frequency oscillations in the Nordic grid using ambient synchrophasor data under the presence of forced oscillations. In *2013 IEEE Grenoble Conference* 1–6 (IEEE, Grenoble, 2013).
61. Cui, Y. et al. Inter-area oscillation statistical analysis of the U.S. Eastern interconnection. *J. Eng.* **2017**, 595–605 (2017).
62. Tchuiseu, E. T., Gomila, D., Brunner, D. & Colet, P. Effects of dynamic-demand-control appliances on the power grid frequency. *Phys. Rev. E* **96**, 022302 (2017).
63. Schröder, M., Zhang, X., Wolter, J. & Timme, M. Dynamic perturbation spreading in networks. *IEEE Trans. Netw. Sci. Eng.* **7**, 1019–1026 (2019).
64. Meyer, P. G., Anvari, M. & Kantz, H. Identifying characteristic time scales in power grid frequency fluctuations with dfa. *Chaos Interdiscip. J. Nonlinear Sci.* **30**, 013130 (2020).
65. Rydin Gorjão, L. MFDFA: multifractal detrended fluctuation analysis in Python. *Zenodo*, <https://zenodo.org/record/3625759> (2020).

Acknowledgements

We express our deepest gratitude to everyone who helped create the database by connecting the EDR at their hotel room, home, or office: Damià Gomila, Malte Schröder, Jan Wohland, Filipe Pereira, André Frazão, Kaur Tuttlberg, Jako Kilter, Hauke Hähne, and Bálint Hartmann. We gratefully acknowledge support from the Federal Ministry of Education and Research (BMBF grant numbers 03SF0472 and 03EK3055), the Helmholtz Association (via the joint initiative Energy System 2050 - A Contribution of the Research Field Energy, the project Uncertainty Quantification - From Data to Reliable Knowledge (UQ) with grant number ZT-I-0029, and the grant number VH-NG-1025), the German Science Foundation (DFG) by a grant toward the Cluster of Excellence Center for Advancing Electronics Dresden (cfaed), and the Scientific Research Projects Coordination Unit of Istanbul University, Project number 32990. This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement number 840825.

Author contributions

L.R.G., R.J., H.M., V.H., G.C.Y., J.K., M.T., C.B., D.W., and B.S. conceived and designed the research. R.J. and H.M. constructed the measurement device and evaluated the experimental data. L.R.G., J.K., and B.S. performed the data analysis and generated the figures. All authors contributed to discussing and interpreting the results and writing the manuscript.

Funding

Open Access funding enabled and organised by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-19732-7>.

Correspondence and requests for materials should be addressed to B.S.

Peer review information *Nature Communications* thanks Pere Colet, Giovanni Filatrella, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Supplementary Information accompanying the manuscript Open data base analysis of scaling and spatio-temporal properties of power grid frequencies

Leonardo Rydin Gorjão,^{1,2} Richard Jumar,³ Heiko Maass,³ Veit Hagenmeyer,³ G.Cigdem Yalcin,⁴
Johannes Kruse,^{1,2} Marc Timme,⁵ Christian Beck,⁶ Dirk Witthaut,^{1,2} and Benjamin Schäfer^{6,*}

¹*Forschungszentrum Jülich, Institute for Energy and Climate Research
- Systems Analysis and Technology Evaluation (IEK-STE), Germany*

²*Institute for Theoretical Physics, University of Cologne, Germany*

³*Karlsruhe Institute of Technology, Institute for Automation and Applied Informatics, Germany*

⁴*Department of Physics, Istanbul University, 34134, Vezneciler, Turkey*

⁵*Chair for Network Dynamics, Center for Advancing Electronics Dresden (cfaed)
and Institute for Theoretical Physics, Technical University of Dresden, Germany*

⁶*School of Mathematical Sciences, Queen Mary University of London, United Kingdom*

Within this Supplementary Information, we give further details on the analysis presented in the main text. In particular, we introduce the abbreviations used to describe the various measurement sites, report the first four statistical moments of all areas. Furthermore, we provide more details on the PCA analysis, including a visualization of the modes and a comparison with the spectrum.

* b.schaefer@qmul.ac.uk

SUPPLEMENTARY NOTE 1

Details on individual measurements: Abbreviations and statistical measures

Abbreviations

The power grid frequency has been recorded in several synchronous areas across Europe and beyond. We introduce the abbreviations used when referring to the measurement sites, e.g. in plots, in Supplementary Table I: We list the town, and country where the measurement was taken and the synchronous area to which this particular device was connected. For example, the measurement taking place in Karlsruhe, Germany is abbreviated as DE (ISO 3166 for Germany) and is connected to the Continental European synchronous area (CE). When discussing the results, we will often refer to the synchronous area and name the measurement site abbreviation in parenthesis, e.g. Continental Europe (DE) is known to be influenced by market dynamics, see also [1]. Unless we are specifically interested in the short time scale, a measurement taken in one location is representative for the whole synchronous area, see also discussion on time-to-bulk in the main text. For areas that are part of a larger country, we first name the country and then specify the area further, e.g. US-UT stands for the United States of America, State Utah, which in turn is part of the Western Interconnection.

Supplementary Table I. Abbreviations of measurement locations and their connection to synchronous areas. For each country, we adopt the ISO 3166 code. The population is extracted from Wikipedia and the sources therein in March 2020.

Abbreviation	Measurement location	Synchronous area	Population
Islands			
IS	Reykjavík, Iceland	Iceland	360,390
FO	Vestmanna, Faroe Islands	Faroe Islands	51,783
ES-GC	Las Palmas de Gran Canaria, Canary Islands, Spain	Gran Canaria	851,231
ES-PM	Palma de Mallorca, Balearic Islands, Spain	Mallorca	896,038
Continental			
DE	Karlsruhe, Germany	Continental Europe (CE)	500,000,000
GB	London, United Kingdom	Great Britain (GB)	66,224,800
EE	Tallin, Estonia	Baltic	6,042,657
SE	Stockholm, Sweden	Nordic	21,180,931
Others			
US-UT	Salt Lake City, Utah, US	Western Interconnection	84,600,000
US-TX	College Station, Texas, US	Texas Interconnection	28,995,881
ZA	Cape Town, South Africa	South Africa	58,775,022
RU	St. Petersburg, Russia	Russia	146,745,098

Statistical measures

Let us systematically investigate the statistical properties of the various areas by computing their mean deviation from the reference μ , as well as their standard deviation σ , skewness β , and kurtosis κ in Supplementary Table II. First, we note that most of the synchronous areas are very close to their reference frequency of 50 Hz (or 60 Hz for US areas) on average. Except for Great Britain (GB) and South Africa (ZA), continental areas display a smaller standard deviation than islanded areas. For GB, we can attribute this large deviations to the very different frequency regulation framework (when compared to Continental Europe), which allows large deviations [2]. The skewness and kurtosis are more difficult to interpret. Some areas, such as Texas (US-TX) and Faroe Islands show a substantial skewness, while other areas, such as Continental Europe (DE) and Iceland (IS) display a large kurtosis $\kappa > 3$, hinting at heavy tails in the distribution.

Kurtosis of aggregated data and increment statistics Let us have a closer look at the kurtosis values. We utilise the kurtosis and its deviation from the Gaussian value of $\kappa^{\text{Gauss}} = 3$ to quantify whether a given distribution displays heavy tails and thereby deviates from a Gaussian distribution. For the aggregated statistics studied above this is relevant as it tells us how likely we will observe extreme deviations, which could lead to a curtailment of demand or a

Supplementary Table II. Statistical properties of different measurement sites. We report the mean μ as the mean deviation from the reference frequency $f - f^{\text{ref}}$, standard deviation σ , skewness β and kurtosis κ of the data.

Area	Mean μ [mHz]	Std. σ [mHz]	Skew. β	Kurtosis κ
Islands				
IS	0.90	55.81	-0.04	7.02
FO	-36.65	133.15	0.27	2.39
ES-GC	-0.60	61.63	0.38	5.00
ES-PM	1.32	70.13	0.06	2.62
Continental				
DE	0.22	18.77	0.08	3.95
GB	-0.06	62.29	0.04	2.42
EE	-0.29	12.44	-0.09	2.95
SE	0.61	48.91	0.19	4.05
Others				
US-UT	-0.73	18.67	-0.04	2.56
US-TX	0.20	17.79	-0.49	2.42
ZA	1.47	93.43	-0.15	2.39
RU	-1.63	16.44	-0.24	2.52

shutdown of generators. For the increments, this is of particular interest for the statistical modelling since a classical Ornstein–Uhlenbeck Process would lead to Gaussian increments [3].

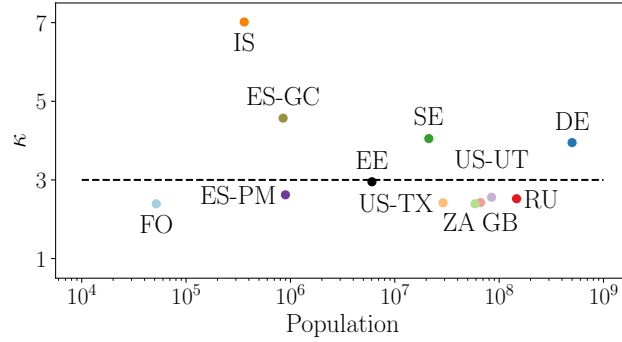
A kurtosis of $\kappa > 3$ is only clearly observed in the aggregated data of IS, ES-GC, DE, and SE. In all cases, the numerous extreme events are a remarkable observation. For such a highly regulated and controlled system as the power grid to deviate so strongly so often demands an explanation. We are confident that the heavy tails at the two continental areas DE and SE are related to extreme deviations at the trading intervals [1, 4, 5]. The tails in the two islands might also be caused by market activities or arise due to a different operation by the TSO, as will be clarified in future work.

The role of the kurtosis changes when moving to increments. A large kurtosis of the increments indicates that the effective noise acting on the system does indeed not follow a Gaussian distribution, as expected in many stochastic processes. Naturally, investigating the increments of an empirical trajectory, we will observe a much more volatile behaviour than if we only observe the trajectory itself. Large jumps will be present and on a short time scale we expect extreme tails in real-world data. Some of these large jumps could be caused by renewable generators, others by fast control mechanisms or inverters. Still it is remarkable that kurtosis values of $\kappa \sim 10^2$, are observed for IS and ES-GC in Fig. 4 of the main text. While the general trend of decreasing kurtosis with increasing time lag is in coherence with previous work on increment analysis [6, 7], further work is necessary to arrive at a full statistical description of frequency increment.

SUPPLEMENTARY NOTE 2

Scaling of fluctuations

Let us investigate whether the fluctuation in terms of regular deviations, measured by the standard deviation σ , or extreme deviations, measured by the kurtosis κ scale with the size of the power system. Intuitively, we would expect that the more generation is present in a given area, the smaller the fluctuations we observe are going to be. As a proxy of the total generation, we use the total population of a synchronous area as this information is easily available for all areas and population and total generation are approximately proportional [8]. As we can see in Supplementary Fig 1, the kurtosis κ is not a simple function of the size of a synchronous area. The occurrence of heavy tails, as measured by the kurtosis κ , depends on dispatch strategies, market regulations [9] and control requirements, which are standardised within the European Network of Transmission System Operators for Electricity (ENTSO-E) [2], leading to very similar values of the kurtosis in most synchronous areas.



Supplementary Figure 1. Extreme fluctuation occurrence do not decrease with increasing grid size. We plot the kurtosis as a measure of heavy tails as a function of the population of a synchronous area.

In contrast, we have seen in the main text that the aggregated noise amplitude ϵ decreases approximately as the square root of the size of a synchronous area. Let us review the derivation of this relation in more detail and discuss alternative approaches to observe the scaling. We follow the arguments presented in [1]: Let us use the standard swing equation [10] to describe the synchronous frequency dynamics at each node i as

$$M_i \dot{\omega}_i(t) = -D_i \omega_i(t) + \epsilon_i \Gamma_i(t) + P_i^m + P_i^e, \quad (1)$$

where $\omega_i = f_i/(2\pi)$ is the nodal angular velocity, M_i is the inertia, D_i is the damping, $\epsilon_i \Gamma_i(t)$ is a noise term and P_i^m and P_i^e are the mechanical power generated or consumed and the transmitted electrical power respectively. Moving to the bulk description, i.e., defining the total inertia $M := \sum_{i=1}^N M_i$ and the bulk angular velocity $\bar{\omega} := \sum_{i=1}^N M_i \omega_i / M$ and assuming a constant damping to inertia ratio $\gamma = D_i / M_i$ [11] as well as balanced electrical and mechanical power on average, i.e., $\sum_{i=1}^N P_i^m = 0$, $\sum_{i=1}^N P_i^e = 0$, we obtain

$$\frac{d}{dt} \bar{\omega}(t) = -\gamma \bar{\omega}(t) + \frac{1}{M} \sum_{i=1}^N \epsilon_i \Gamma_i(t). \quad (2)$$

The equation from the main text is re-obtained by identifying ΔP as $\sum_{i=1}^N \epsilon_i \Gamma_i(t)$. If we assume that the noise $\Gamma_i(t)$ at each node is approximately Gaussian with zero mean and standard deviation 1 (the amplitude is included in ϵ_i), we can formulate the Fokker–Planck equation [3] of this Ornstein–Uhlenbeck process (2) as

$$\frac{\partial p}{\partial t} = \gamma \frac{\partial}{\partial \bar{\omega}} (\bar{\omega} p) + \frac{1}{2} \frac{1}{M^2} \left[\sum_{i=1}^N \epsilon_i^2 \right] \frac{\partial^2 p}{\partial \bar{\omega}^2}. \quad (3)$$

The resulting probability density function $p(\bar{\omega})$ is a normal distribution with mean 0 and standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^N \epsilon_i^2}{\gamma M^2}}. \quad (4)$$

As an additional simplification, let us assume identical noise $\epsilon_i = \xi$ and identical inertia $M_i = m$ at all nodes, i.e., $M = Nm$. Then, the standard deviation is given as

$$\sigma = \sqrt{\frac{N\xi^2}{\gamma m^2 N^2}} = \sqrt{\frac{\xi^2}{\gamma m^2}} \frac{1}{\sqrt{N}}. \quad (5)$$

This means the standard deviation decays approximately as $\sigma \sim 1/\sqrt{N}$. An important assumption when comparing areas with Supplementary Eq. (5) is that we assume similar noise ξ , damping γ , and inertia m . While the inertia per machine will likely be similar in the different areas, the effective damping depends on the control applied and the noise on the mix of generators (nuclear, hydro, coal, wind, solar, ...) and the nature of the demand fluctuations.

To take these additional external factors into account when comparing the empirical data of the various synchronous areas, we do not compare the standard deviation several areas but their aggregated noise amplitudes ϵ . To retrieve the aggregated noise amplitude ϵ we employ a non-parametric Nadaraya–Watson estimator to extract the Kramers–Moyal moments of the underlying stochastic dynamics. For a timeseries $x(t)$, the n th Kramers–Moyal moment can be extracted via

$$M_n(x, t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \langle (x'(t + \Delta t) - x'(t))^n | x'(t) = x \rangle, \quad (6)$$

where the averaging process is made more precise by implementing a kernel-density function $K_h(\cdot) = h^{-1}K(\cdot/h)$ with a bandwidth h . The Nadaraya–Watson estimator $W_h(x)$, at point i , is given by [12]

$$W_h(x)_i = \frac{K_h(x - x'_i)}{\sum_{j=1}^S K_h(x - x'_j)}, \quad (7)$$

where we take $K_h(x)$ to be an Epanechnikov function, compact in $\mathbb{R}_{[-1,1]}$, and S is the number of data points. The aggregated noise amplitude ϵ is retrieved studying the second Kramers–Moyal moment, and employing the aforementioned estimator, resulting in

$$M_2(x) = \frac{1}{S\Delta t} \sum_{i=1}^S W_h(x)_i [x'(t + \Delta t) - x'(t)]_i^2 = \epsilon^2, \quad (8)$$

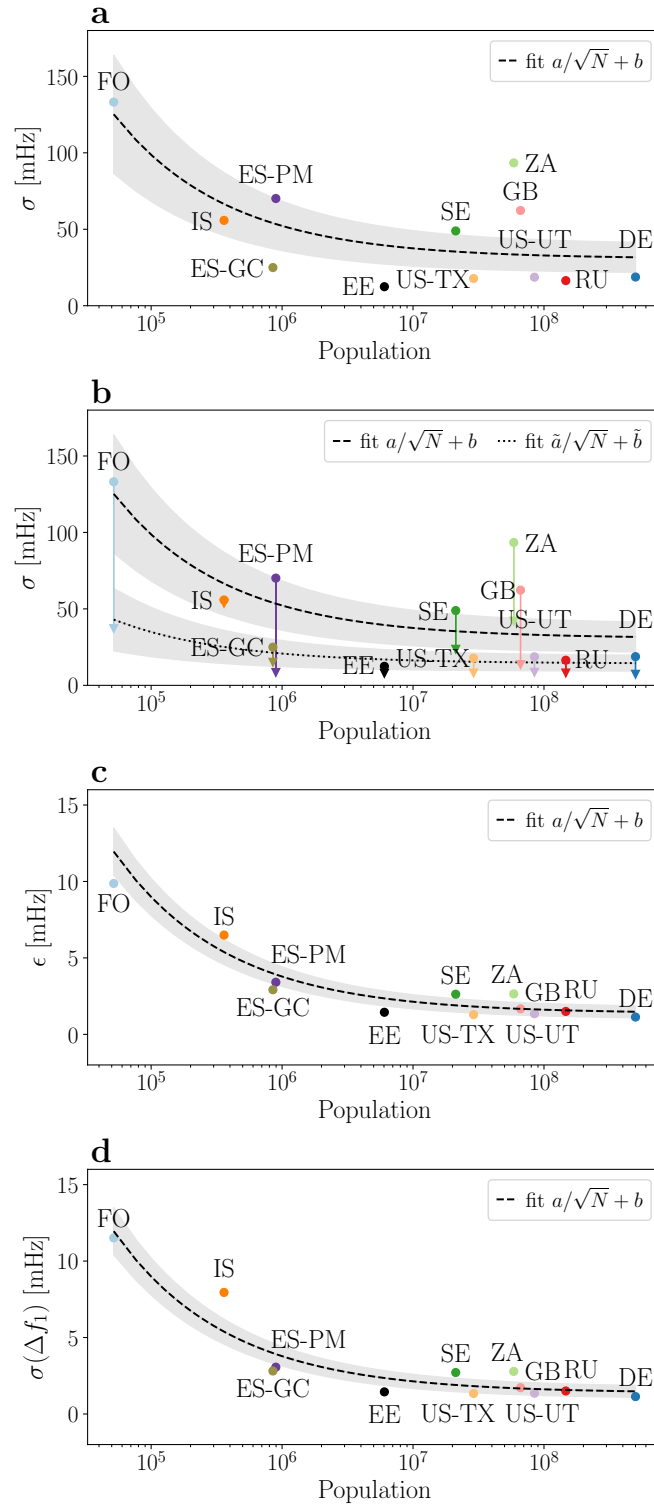
with $\Delta t = 1s$. All Kramers–Moyal moments are easily retrievable, see Refs. [13, 14]. The so extracted aggregated noise amplitude ϵ will closely follow the empirical Fokker–Planck equation and should therefore resemble a $\epsilon \sim 1/\sqrt{N}$ decay, as any deterministic effects are filtered out.

Finally, while we expect the noise to decay with the size, it is well-justified to add a constant to our previously derived expression (5), modifying it to

$$\epsilon \sim \sigma = \frac{a}{\sqrt{N}} + b. \quad (9)$$

We add the constant b to take the effect of deadbands into account. All synchronous areas have deadbands [10], i.e. frequency ranges for which there is no (primary) control active and the frequency dynamics evolves freely. This means for a certain range $|\omega| < \omega_{\text{deadband}}$ the damping to inertia ratio γ , which explicitly includes primary control, is much smaller and the frequency randomly evolves and contributes a minimum noise contribution b .

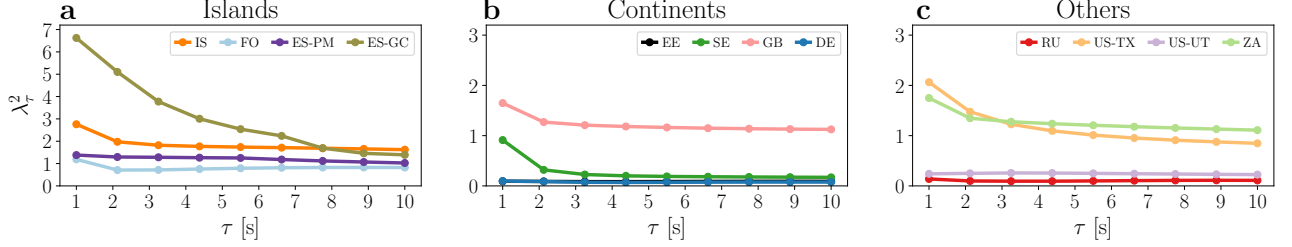
In addition to the standard deviation σ or the aggregated noise amplitude ϵ , we might also consider using a Gaussian kernel to detrend the data and only analyse the standard deviation of the detrended data. Finally, we could also consider using the standard deviation of the increments (at the smallest available time lag $\tau = 1$) as a proxy of the actual noise. We compare these different approaches in Supplementary Fig. 2: Moving from top to bottom: Taking the unfiltered standard deviation of the data (Supplementary Fig. 2 a), does give a rough trend but areas as South Africa (ZA) or Great Britain (GB) have a much larger standard deviation than we would expect from the scaling law. When we introduce the de-trending, the overall standard deviation drops (Supplementary Fig. 2 b) but the problems with GB and ZA persist. Likely this is caused by control regulations that allow larger deviations than they are allowed for example in Continental Europe. Next, we extract the aggregated noise amplitude ϵ (Supplementary Fig. 2 c) and observe a decay, well-approximating the conjectured scaling law. Finally, we note that using the standard deviation of the increments at lag $\tau = 1s$ is almost identical to the aggregated noise amplitude ϵ (Supplementary Fig. 2 d). Comparing all four plots, in particular in terms of the standard deviation of the best fit (shaded area), leads to the conclusion that panels c and d are the best descriptors. Both the extracted noise ϵ as well as the increments on the one second scale Δf_1 are good proxies of the diffusive forces acting on the synchronous area.



Supplementary Figure 2. Average fluctuations decrease with increasing grid size with increments Δf_1 and aggregated noise amplitude ϵ as the best descriptors. We plot the standard deviation (a), filtered standard deviation (b), the aggregated noise amplitude (c) and the standard deviation of the increments at time lag $\tau = 1s$ (d). The shaded area gives the standard deviation of the fit.

SUPPLEMENTARY NOTE 3

Castaing's model for increments



Supplementary Figure 3. Increment analysis reveals non-Gaussian characteristics, dominantly in islands. We plot the Castaing parameter λ_τ^2 , given by Supplementary Eq. (14), for the different examined power grid frequency recordings. We observe a non-vanishing intermittency in Gran Canaria (ES-GC), Iceland (IS), Faroe Islands (FO), Mallorca (ES-PM), Britain (GB), Texas (US-TX), and South Africa (ZA). In contrast, the increments' distribution of the Baltic (EE), Continental Europe (DE), Nordic (SE), Russia (RU) synchronous areas and the Western Interconnection (US-UT) approach a Gaussian distribution.

In the main text we quantified deviations from Gaussianity of the increment distributions by the use of the excess kurtosis $\kappa - 3$, which should decay to zero if the distributions under consideration are Gaussian. Here, we offer a more theoretical view by using Castaing's model. This model describes the deviations of the increments' distribution from a Gaussian distribution [15–17] and has already been applied to frequency analysis [6, 7]. The increments Δf_τ with the lag τ have a non-homogenous scaling, which leads to distributions with high kurtosis, and sometimes non-zero skewness [18]. The rationale is that the process is a superposition of several subset processes with distinct scales, similar to superstatistics [19, 20]. Specifically, Castaing's model is a special case of log-normal superstatistics applied to increments. The probability density function (PDF) of the increments $p(\Delta f_\tau, \sigma_\tau)$ is a function of the widths σ_τ , given by

$$p(\Delta f_\tau, \tau) = \int_{-\infty}^{\infty} L_\lambda(\sigma_\tau) p_0(\Delta f_\tau, \sigma_\tau) d \ln \sigma_\tau, \quad (10)$$

where the underlying subset processes are assumed to have a Gaussian distribution $p_0(\Delta f_\tau, \sigma_\tau) \sim \mathcal{N}(0, \sigma_\tau)$ of some variance σ and $L_\lambda(\sigma_\tau)$ accounts thus for the scales of each superposition. This scale function $L_\lambda(\sigma_\tau)$ is conjectured to be log-normally distributed

$$L_\lambda(\sigma_\tau) = \frac{1}{\lambda_\tau \sqrt{2\pi}} \exp \left[-\frac{\ln^2(\sigma_\tau/\sigma_0)}{2\lambda_\tau^2} \right], \quad (11)$$

with λ_τ^2 being the Castaing parameter. For the case $\lambda_\tau^2 \rightarrow 0$, the distribution $L_\lambda(\sigma_\tau)$ approached a δ -distribution, and the increments Δf_τ are purely Gaussian distributed. As λ_τ^2 increases, the convolution includes more scales and the tails of the PDF of the increments enlarge. Inserting Supplementary Eq. (11) into Supplementary Eq. (10) yields the explicit PDF of the increments as

$$p(\Delta f_\tau, \tau) = \frac{1}{\lambda_\tau 2\pi} \int_0^\infty \frac{d\sigma_\tau}{\sigma_\tau^2} \exp \left[-\frac{\Delta f_\tau^2}{2\sigma_\tau^2} - \frac{\ln^2(\sigma_\tau/\sigma_0)}{2\lambda_\tau^2} \right]. \quad (12)$$

The increments intermittency behaviour is thus solely described by the Castaing parameter λ_τ^2 . Multiplying both sides of Supplementary Eq. (12) by Δf_τ^2 and integrating over $[-\infty, \infty]$, we find

$$\sigma_0^2 = \langle \Delta f_\tau^2 \rangle \exp[-2\lambda_\tau^2]. \quad (13)$$

One can now recover the Castaing parameter λ_τ^2 by extracting the fourth-order statistical moment, i.e., the kurtosis $\kappa_{\Delta f}(\tau)$, of the PDFs of Δf_τ as function of the lag τ . The Castaing parameter λ_τ^2 results in

$$\lambda_\tau^2 = \ln \left(\frac{\kappa_{\Delta f}(\tau)}{3} \right), \quad (14)$$

i.e., for a Gaussian distribution with $\kappa_{\Delta f}(\tau) = 3$ the Castaing parameter decays to 0.

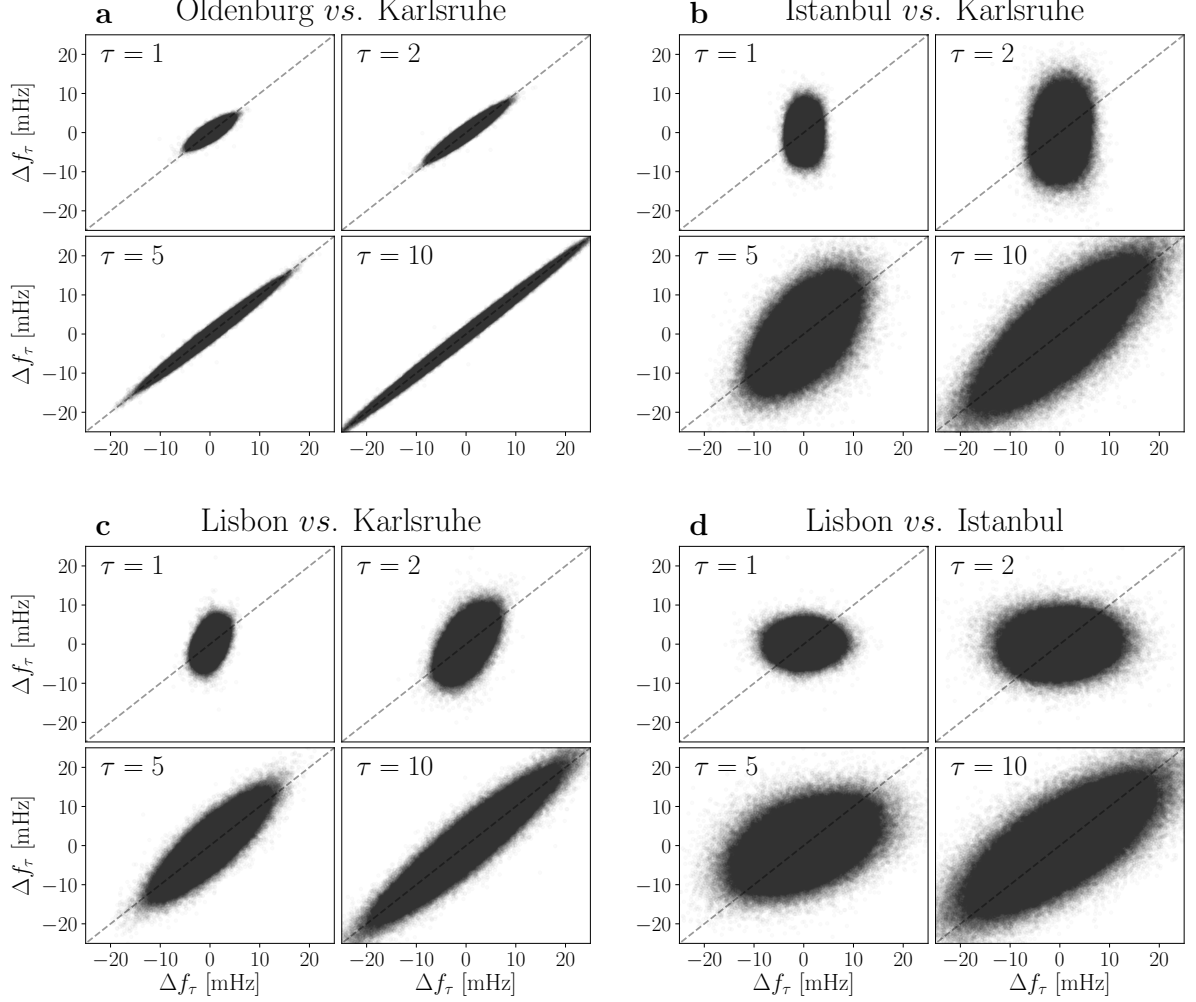
To extract the Castaing parameter, we compute the increment statistics Δf_τ of a certain timeseries f_τ and calculate the kurtosis κ_τ of the PDF of the increments. Then, we take the normalised logarithm as in Supplementary Eq. (14) and do so over the desired range of increments time-lag τ .

Computing the Castaing parameter λ_τ^2 for our data, we observe very heterogeneous results between the various synchronous areas, see Supplementary Fig. 3. In some areas, the intermittent behaviour of the increments Δf_τ is subdued and the overall distribution approaches a Gaussian distribution (in EE, DE, SE, RUS, and US-TX), i.e., the Castaing parameter λ_τ^2 approaches 0. On the other hand, all islands display large and non-vanishing intermittent behaviour, as well as GB, US-TX, and ZA. Iceland (IS) but particularly Gran Canaria (ES-GC) shows impressive deviations from Gaussianity that require detailed modelling in the future.

With Castaing's model and parameter we used an alternative approach to quantify heavy tails, instead of purely using the kurtosis κ . A further advantage of Castaing's approach is that it allows us to model the increments as superimposed distributions, complementary to superstatistics of the aggregated frequency distributions [1]. An explicit increment model will be left for future work.

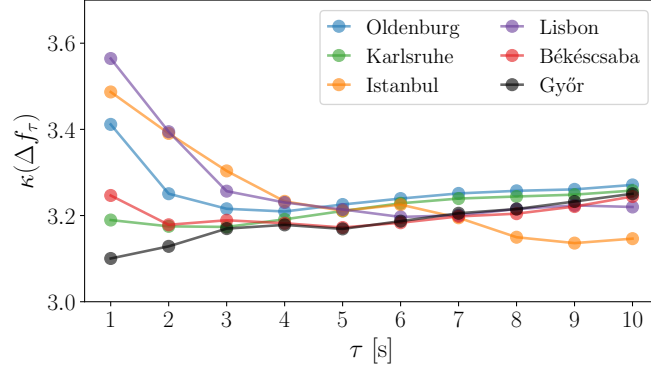
SUPPLEMENTARY NOTE 4

Further increment analysis



Supplementary Figure 4. Long increment lags lead to increased correlations. We repeat the increment analysis from the main text but with larger lags $1 \text{ s} \leq \tau \leq 10 \text{ s}$. Note that each of the 2×2 subpanels is now using four different lags τ but the overall figure still follows the same arrangement as in the main text.

We complement the increment analysis from the main text by considering larger time lags $\tau > 1 \text{ s}$. In particular, we compute the increments [6, 21] $\Delta f_\tau = f(t+\tau) - f(t)$ for the four measurement sites in Continental Europe: Karlsruhe, Oldenburg, Istanbul and Lisbon. For the pairs investigated in the main text, we consider a lag $\tau = 1, 2, 5, 10$ seconds in Supplementary Fig. 4. Since the scatter plots report increments at the same time t , points on the diagonal indicate large correlations, while circles or ellipses aligned with one axis indicate no correlations. The results for Oldenburg vs. Karlsruhe are almost independent of the specific time lag τ . The magnitude of the increments increases for increasing time lag τ but almost all values follow the diagonal, indicating a high correlation on the full time scale for 1 to 10 seconds. In contrast, the increment plots involving Istanbul and Lisbon change much more with increasing lag τ as their increments on the time scale of 1 second are almost completely uncorrelated but at 5 to 10 seconds, an increasing number of points follow the diagonal, i.e., fluctuation events become correlated. Similar to the detrended fluctuation analysis (DFA) from the main text, we again observe that short time scales are independent, while we approximate a bulk description for longer time scales. This observation is also consistent with claims found e.g in [22] that high frequency fluctuations do not penetrate the grid over long distances but lower frequency fluctuations do.



Supplementary Figure 5. The kurtosis κ of the increment statistics Δf_τ decreases with increasing lag τ . We plot the kurtosis κ of each recording site as a function of the increment lag τ .

Finally, we may further investigate how the increment distributions look like, in particular with respect to their large deviations, i.e., their heavy tails measured by the kurtosis of the increment distributions. Computing the kurtosis κ of the increment statistics Δf_τ at different lags τ , shows that the deviations from the Gaussian ($\kappa^{\text{Gaussian}} = 3$) decrease *on average*. While the kurtosis shows a small increase in Győr with increasing time lag, the kurtosis at Istanbul and Lisbon is substantially reduced, see Supplementary Fig. 5. This finding is consistent with [6], where the authors also found that long lags lead to approximately Gaussian increment statistics. Here, we go further in that we observe spatial differences already at a time resolution of 1 second, in particular for locations far away from Karlsruhe.

SUPPLEMENTARY NOTE 5

Details on Detrended Fluctuation Analysis (DFA)

Detrended Fluctuation Analysis (DFA) [23, 24] studies the fluctuation of a given process by considering increasing segments of the timeseries. Take a timeseries $X(t)$ with N elements X_i , $i = 1, 2, \dots, N$. Obtain the detrended profile of the process by defining

$$Y_i = \sum_{k=1}^i (X_k - \langle X \rangle), \text{ for } i = 1, 2, \dots, N,$$

i.e., the cumulative sum of X subtracting the mean $\langle X \rangle$ of the data. Section the data into smaller non-overlapping segments of length s , obtaining therefore $N_s = \text{int}(N/s)$ segments. Given the total length of the data is not always a multiple of the segment's length s , discard the last points of the data. Consider the same data, apply the same procedure, but this time discard instead the first points of the data. One has now $2N_s$ segments. To each of these segments fit a polynomial y_v of order m and calculate the variance of the difference of the data to the polynomial fit

$$F^2(v, s) = \frac{1}{s} \sum_{i=1}^s [Y_{(v-1)s, i} - y_{v, i}]^2, \text{ for } v = 1, 2, \dots, N_s,$$

where $y_{v, i}$ is the polynomial fitting for the segment i of length v . One also has the freedom to choose the order of the polynomial fitting. This gives rise to the denotes DFA1, DFA2, \dots , for the orders chosen. Notice $F^2(v, s)$ is a function of each variance of each v -segment of data and of the different s -length segments chosen. One can define the fluctuation function $F^2(s)$ by averaging each row of segments of size s

$$F^2(s) = \frac{1}{N_s} \left\{ \sum_{v=1}^{N_s} F^2(v, s) \right\}^{1/2}.$$

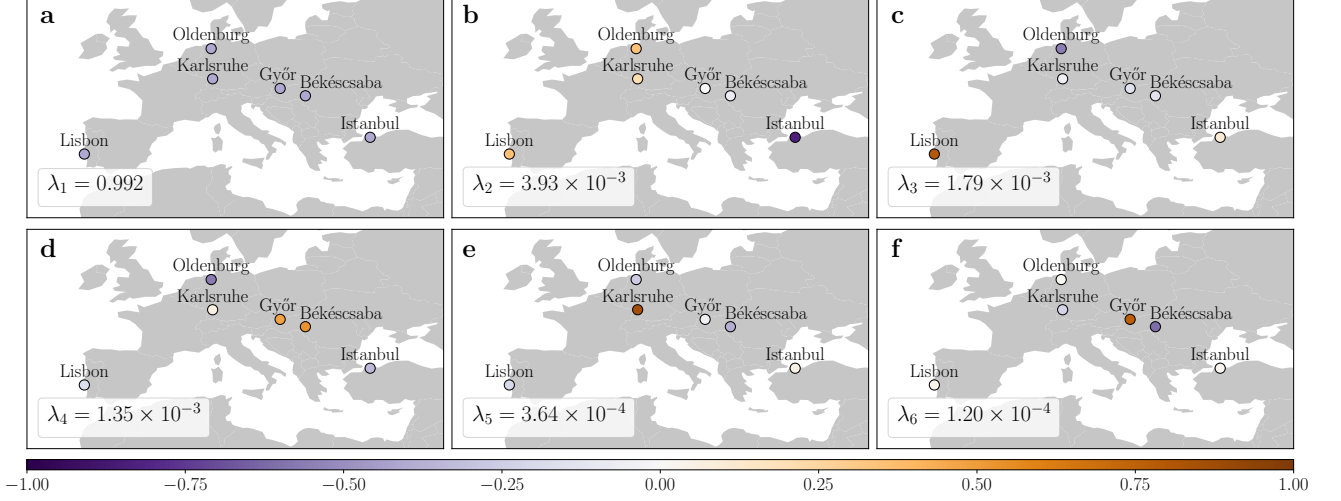
The inherent scaling properties of the data, if the data displays power-law correlations, can now be studied in a log-log plot of $F^2(s)$ versus s . Herein the scaling of the data obeys a power-law with exponent h as

$$F^2(s) \sim s^h,$$

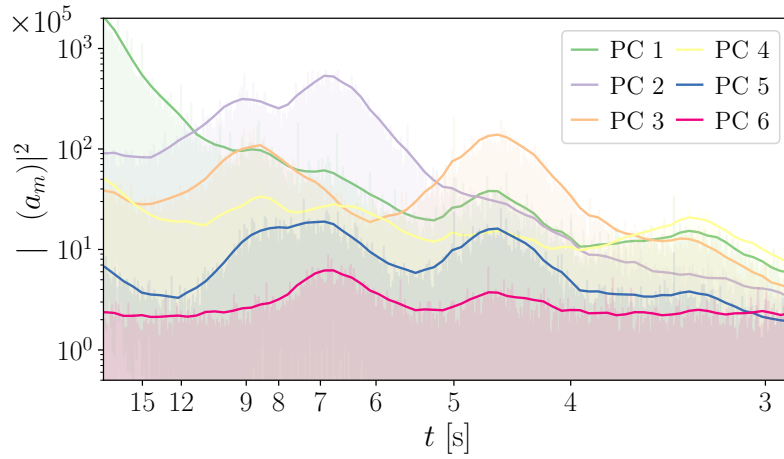
where h is the *self-similarity* exponent (which may be multifractal) and relates directly to the Hurst index. The self-similarity exponent h is calculated by finding the slope of this curve in the log-log plots. For a detailed explanation of DFA, see [25]. The analysis implemented here is based on [26].

SUPPLEMENTARY NOTE 6

Principal Component Analysis



Supplementary Figure 6. The principal components reveal the spatial structure of inter-area modes. The maps show the colour-coded entries of the normalised principal components \mathbf{f}_m . The spatial structure corresponds to synchronous behaviour (a), East-West oscillations (b), North-South oscillations (c), Coastal-inland fluctuations (d-e), and intra-Hungarian oscillations (f). The synchronous component describes the frequency bulk behaviour and thus explains already $\lambda_1 = 99.2\%$ of the total variance.



Supplementary Figure 7. The inter-area modes oscillate with a period between 3 and 9 seconds. The figure displays squared Fourier amplitudes $|\mathcal{F}(a_m(t))|^2$ of the mode amplitudes $a_m(t)$. The largest inter-area oscillations occur with a period of $t = 7$ s and $t = 4.5$ s, which corresponds well to the results in [27]. In the main text, we only discussed the first three modes as the most dominant ones.

We apply a principal component analysis (PCA) to our multivariate frequency time series in order to extract the dominant modes of inter-area oscillations. Generally, a PCA identifies the orthogonal linear subspaces (principal components) that maximise the projected variance of the data [28]. Let $\mathbf{f}(t)$ be the vector containing the frequency recordings from all six locations within Europe. The principal components are then given by the eigenvectors $\mathbf{f}_m(t)$ of the data covariance matrix. Based on the principal components, we can decompose the centred frequency recordings

into uncorrelated time series $a_m(t)$:

$$\mathbf{f}(t) = \langle \mathbf{f} \rangle + \sum_{m=1}^N a_m(t) \mathbf{f}_m.$$

The time series $a_m(t)$ describe the amplitude of the frequency recordings $\mathbf{f}(t)$ projected onto the m -th principal component. Their variance, i.e., the projected variance, is given by the eigenvalue $\tilde{\lambda}_m$. By rescaling this variance, we obtain the share of total variance explained by the m -th component:

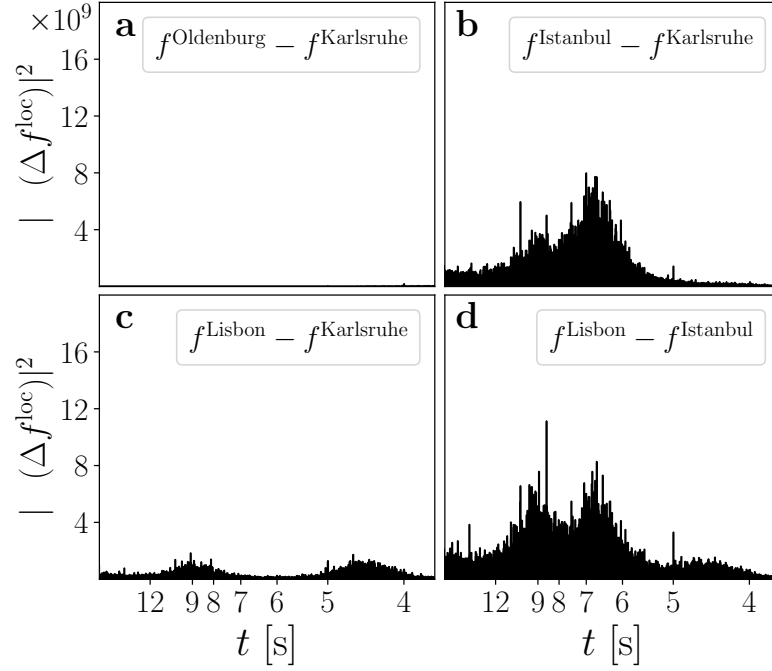
$$\lambda_m = \frac{\tilde{\lambda}_m}{\sum_m \tilde{\lambda}_m}.$$

We interpret the principal components (PCs) of $\mathbf{f}(t)$ as spatial inter-area modes and the time series $a_m(t)$ as their amplitudes. Supplementary Fig. 6 displays all six modes obtained from our synchronised measurements in Continental Europe. The first mode (PC1) represents a synchronous frequency oscillation at all locations and thus corresponds to the bulk behaviour of the frequency. The other modes (PC2-PC6) display asynchronous spatial patterns and thus represent inter-area oscillations. Due to their small amplitude, these modes only explain a low portion of the variance ($\lambda_2, \dots, \lambda_6 < 0.4\%$), while the bulk mode already covers $\lambda_1 \approx 99.2\%$. However, PC2 to PC6 uncover the dominant spatial structure of inter-area modes across Continental Europe. Let us analyse these modes based on their visualisation in Supplementary Fig. 6 in more detail: PC2 contains one coherent area in Western Europe that oscillates in phase opposition to Istanbul. In PC3 Lisbon and Istanbul swing in opposition to the Northern measurement locations. PC2 thus resembles an East-West dipole, while PC3 is similar to a North-South dipole. The other modes correspond to a Coast-Inland fluctuation (PC4 and PC5) and an intra-Hungarian oscillation (PC6).

To reveal the oscillation period of these modes, we analyse the power spectral density (PSD) of their amplitudes $a_m(t)$ (Supplementary Fig. 7). The East-West dipole (PC2) exhibits a main period length of $t \approx 7$ s and a smaller contribution at $t \approx 9$ s. PC3 mainly oscillates with a period of $t \approx 4.5$ s, but there is also a distinct peak at $t \approx 9$ s, which also appears in PC2 and PC4. PC5 and PC6 mainly exhibit oscillations with a period of $t \approx 7$ s and $t \approx 4.5$ s. The PSD of bilateral differences between single measurement locations reveals the same main period lengths (Supplementary Fig. 8). Interestingly, no inter-area oscillations can be observed between Karlsruhe and Oldenburg, which could indicate well-balanced power within Germany or could be caused by the limited spatial resolution of the available 6 modes. Overall, we identify three main oscillations periods of inter-area modes with $t \approx 9$ s, $t \approx 7$ s, and $t \approx 4.5$ s, which is consistent with the typical frequency of inter-area modes [29].

The comparison to a more detailed study suggests that our principal components probably relate to overlaps of different global inter-area modes. A first indicator for this conclusion is the occurrence of multiple substantial peaks in our PSD in Supplementary Fig. 7 (e.g. for PC3). The detailed analysis in [27] revealed four dominant inter-area modes in Continental Europe with distinct period lengths. The authors describe a mode G1, which resembles a North-South dipole and oscillates with $t = 5$ s. Furthermore, they present mode T1, which represents an East-West dipole oscillating at $t = 6.7$ s, and similar mode G2 with $t = 3.3$ s. In mode G3, Eastern Europe, Spain, and Portugal swing in phase opposition to Central Europe with $t = 2$ s. Comparing this to our results, we conclude that PC2 corresponds well to mode T1, while PC3 is similar to mode G1. However, PC3 contains another substantial oscillation with $t = 9$ s. This could be an effect of aliasing due to our low Nyquist-frequency of 0.5 Hz. Modes with period lengths below 2 s thus re-appear in our PSD at multiples of their oscillation period. The mode G3 could thus correspond to the peak at $t \approx 9$ s in Supplementary Fig. 7. Finally, PC4 contains the mode G2 among others, while PC5 and PC6 both contain the period lengths of G1 and T1. The PCA modes are thus very similar to the results in [27], but they do not isolate inter-area modes with single distinct periods.

The overlap of different (linear) oscillation modes in our PCA results can have different reasons. Our low spatial resolution could make it impossible to retain the spatial modes identified in an earlier study [27]. On the other hand, a linear separation of the inter-area modes through a PCA and a Fourier decomposition could generally fail due to the non-linearity of power system dynamics. Finally, our time resolution of one second could lead to aliasing and the incorrect reconstruction of inter-area modes. In the future, a higher spatial and temporal resolution of frequency recordings would be necessary to further investigate these effects.



Supplementary Figure 8. Inter-area oscillations between different grid sites. We perform a spectral analysis of the frequency differences by computing the spectrum of the frequency difference Δf^{loc} between two sides, e.g. in (e): $\Delta f^{\text{loc}} = f^{\text{Oldenburg}} - f^{\text{Karlsruhe}}$.

SUPPLEMENTARY REFERENCES

-
- [1] Schäfer, B., Beck, C., Aihara, K., Witthaut, D. & Timme, M. Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics. *Nature Energy* **3**, 119–126 (2018).
 - [2] ENTSO-E. Network code on requirements for grid connection applicable to all generators (rfg). <https://www.entsoe.eu/major-projects/network-code-development/requirements-for-generators/> (2013).
 - [3] Gardiner, C. W. *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences* (Springer, 1985).
 - [4] Rydin Gorjão, L. *et al.* Data-driven model of the power-grid frequency dynamics. *IEEE Access* **8**, 43082–43097 (2020).
 - [5] Schäfer, B., Timme, M. & Witthaut, D. Isolating the impact of trading on grid frequency fluctuations. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 1–5 (IEEE, 2018).
 - [6] Hähne, H., Schottler, J., Wächter, M., Peinke, J. & Kamps, O. The footprint of atmospheric turbulence in power grid frequency measurements. *Europhysics Letters* **121**, 30001 (2018).
 - [7] Hähne, H., Schmietendorf, K., Tamrakar, S., Peinke, J. & Kettemann, S. Propagation of wind-power-induced fluctuations in power grids. *Physical Review E* **99**, 050301 (2019).
 - [8] ENTSO-E. Statistical factsheet 2018. https://docstore.entsoe.eu/Documents/Publications/Statistics/Factsheet/entsoe_sfs2018_web.pdf (2018).
 - [9] Weißbach, T. & Welfonder, E. High Frequency Deviations within the European Power System—Origins and Proposals for Improvement. *VGB powertech* **89**, 26 (2009).
 - [10] Machowski, J., Bialek, J. & Bumby, J. *Power System Dynamics: Stability and Control* (John Wiley & Sons, Chichester, 2011).
 - [11] Weixelbraun, M., Renner, H., Schmaranz, R. & Marketz, M. Dynamic simulation of a 110-kv-network during grid restoration and in islanded operation. In *20th International Conference and Exhibition on Electricity Distribution-Part 1, 2009*, 1–4 (IET, 2009).
 - [12] Lamouroux, D. & Lehnertz, K. Kernel-based regression of drift and diffusion coefficients of stochastic processes. *Physics Letters A* **373**, 3507–3512 (2009).
 - [13] Rydin Gorjão, L. & Meirinhos, F. kramersmoyal: Kramers–Moyal coefficients for stochastic processes. *Journal of Open Source Software* **4**, 1693 (2019).
 - [14] Rinn, P., Lind, P. G., Wächter, M. & Peinke, J. The Langevin approach: An R package for modeling Markov processes.

- Journal of Open Research Software* **4**, e34 (2016).
- [15] Castaing, B., Gagne, Y. & Hopfinger, E. Velocity probability density functions of high Reynolds number turbulence. *Physica D: Nonlinear Phenomena* **46**, 177–200 (1990).
 - [16] Castaing, B. Scalar intermittency in the variational theory of turbulence. *Physica D: Nonlinear Phenomena* **73**, 31–37 (1994).
 - [17] Castaing, B. The temperature of turbulent flows. *Journal de Physique II* **6**, 105–114 (1996).
 - [18] Tabar, M. R. R. *The Friedrich–Peinke Approach to Reconstruction of Dynamical Equation for Time Series: Complexity in View of Stochastic Processes*, 143–164 (Springer International Publishing, Cham, 2019).
 - [19] Beck, C. & Cohen, E. G. D. Superstatistics. *Physica A* **322**, 267–275 (2003).
 - [20] Beck, C., Cohen, E. G. D. & Swinney, H. L. From time series to superstatistics. *Physical Review E* **72**, 056133 (2005).
 - [21] Anvari, M. *et al.* Short term fluctuations of wind and solar power systems. *New Journal of Physics* **18**, 063027 (2016).
 - [22] Zhang, X., Hallerberg, S., Matthiae, M., Witthaut, D. & Timme, M. Fluctuation-induced distributed resonances in oscillatory networks. *Science Advances* **5**, eaav1027 (2019).
 - [23] Peng, C.-K. *et al.* Mosaic organization of DNA nucleotides. *Physical Review E* **49**, 1685–1689 (1994).
 - [24] Kantelhardt, J. W. *et al.* Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A* **316**, 87–114 (2002).
 - [25] Ihlen, E. Introduction to Multifractal Detrended Fluctuation Analysis in Matlab. *Frontiers in Physiology* **3**, 141 (2012).
 - [26] Rydin Gorjão, L. MFDFA: Multifractal Detrended Fluctuation Analysis in Python. <https://zenodo.org/record/3625759> (2020).
 - [27] Grebe, E., Kabouris, J., Lopez Barba, S., Sattinger, W. & Winter, W. Low frequency oscillations in the interconnected system of Continental Europe. In *IEEE PES General Meeting*, 1–7 (IEEE, Minneapolis, MN, 2010).
 - [28] Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2007), 1 edn.
 - [29] Klein, M., Rogers, G. J. & Kundur, P. A fundamental study of inter-area oscillations in power systems. *IEEE Transactions on Power Systems* **6**, 914–921 (1991).

2.2.2 Publication #5

L. Rydin Gorjão, L. Vanfretti, D. Witthaut, C. Beck and B. Schäfer, under the working title *Phase and amplitude synchronisation in power-grid frequency fluctuations*, Ref. [5].

Status: in preparation

Phase and amplitude synchronisation in power-grid frequency fluctuations

Leonardo Rydin Gorjão,^{1,2,*} Luigi Vanfretti,^{3,†} Dirk Witthaut,^{1,2,‡} Christian Beck,^{4,§} and Benjamin Schäfer^{4,¶}

¹*Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany*

²*Forschungszentrum Jülich, Institute for Energy and Climate Research - Systems Analysis and Technology Evaluation (IEK-STE), 52428 Jülich, Germany*

³*Electrical, Computer, and Systems Engineering,*

Rensselaer Polytechnic Institute, 8024 Troy, New York, United States

⁴*School of Mathematical Sciences, Queen Mary University of London, United Kingdom*

Monitoring and modelling the power grid frequency is key to ensuring stability in the electrical power system. Many tools exist to investigate the detailed deterministic dynamics and especially the bulk behaviour of frequency. However, far less attention has been paid to its stochastic properties, and no coherent framework connects both short-time scale fluctuations and bulk behaviour. Moreover, commonly assumed uncorrelated stochastic noise is predominantly employed in modelling in energy systems. In this publication, we examine the stochastic properties of six synchronous power-grid frequency recordings with high-temporal resolution of the Nordic Grid from September 2013, focusing on the increments of the frequency recordings. We show that these increments follow non-Gaussian statistics and display spatial and temporal correlations. Furthermore, we report two different physical synchronisation phenomena: a very short timescale phase synchronisation (< 2 s) followed by a slightly larger timescale amplitude synchronisation (2 s–5 s). Overall, these results provide guidance how to model fluctuations in power systems.

I. INTRODUCTION

The power-grid frequency is a key indicator of the stability of electric power systems. It weighs in the balance of power generation and consumption as well as the operation of each element of the grid and thus serves as the main observable in automatic generation control [1]. In steady operation, the frequency is the same throughout the grid and all generators have a fixed phase difference that essentially determines the real power flows. Yet, perfect phase locking is only an approximation to the operation of real power grids subject to numerous external perturbations. For instance, local perturbations, such as the loss of a generating unit, can cause inter-area oscillations corresponding to the normal modes of the grid around a steady state [2]. Stronger perturbations can even lead to the complete loss of synchrony between different parts of the power grid, eventually leading to blackouts [3]. This article focuses on mild deviations caused by ambient perturbations and the ability of the grid to relax to synchrony afterwards.

Frequency dynamics and synchronisation are essential aspects of power system operation and thus intensively studied in the literature. The ongoing energy transition strengthened the interest in these topics, as synchronous generators are replaced by inverter-based power sources [4, 5]. For instance, recent years saw an enormous progress in the mathematical theory of synchronisation [6], leading to the derivation of a variety of rigorous

stability conditions [7, 8]. An important actual research topic is the dynamics and design of inverter-based power grids lacking the inertia provided by large synchronous machines [9, 10]. Another key approach is the development of detailed simulation models to study frequency dynamics and synchronisation for actual grid layouts and contingency situations [11, 12]. Power hardware in the loop then allows to investigate and test actual equipment coupled with real-time simulations [13, 14]. A common scenario in such simulation-based studies is the dynamics after a sudden large perturbation modelling a fault (see, e.g. Ref. [4]). Does the grid remain stable and how does it take to relax to a steady synchronous operation? The current manuscript adopts a very different approach to power-grid synchronisation focusing on the analysis of synchronous frequency measurements in the presence of ambient noise. We will extract the essential scales of synchronisation in space and time from frequency time series in a model-free way.

Non-linear time series analysis has been applied to investigate various aspects of power-grid systems, including stochastic analysis [15], particularly with the application of the Hilbert–Huang transform (empirical mode decomposition) [16], wavelet-based analysis [17], or power spectral density. The spatio-temporal dynamics of the power system have been studied in more detail, e.g. by revealing inter- and intra-area oscillations and eigenfrequencies. In the Nordic Grid in particular, eigenfrequencies [18] and eigenmodes [19] have been estimated and inter-area oscillations [20] as well as power oscillation damping [21] have been observed. Nevertheless, stochastic elements, more commonly denoted as ambient noise [22], are not in the focus of most power system studies. Only some studies include noise, e.g. when investigating power generation [23]. Some recent papers have provided a detailed stochastic analysis and modelling of

* l.rydin.gorjao@fz-juelich.de

† luigi.vanfretti@gmail.com

‡ d.witthaut@fz-juelich.de

§ c.beck@qmul.ac.uk

¶ b.schaefer@qmul.ac.uk

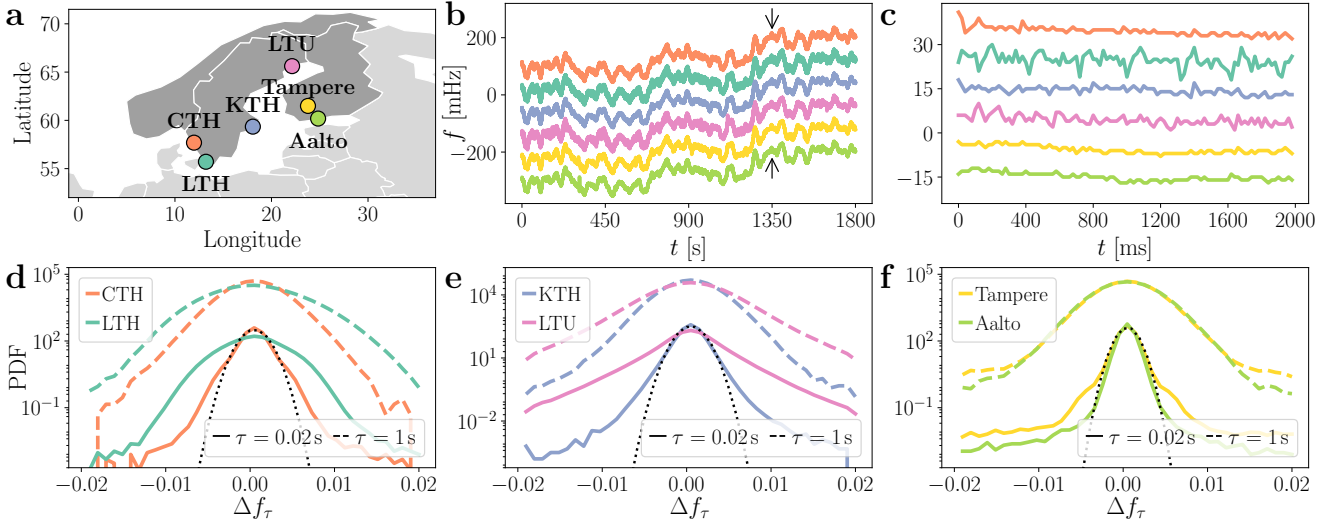


FIG. 1. Six power-grid frequency recordings in the Nordic Grid from 2013, showing very distinct increment statistics. (a) Approximate locations of the recordings across the Nordic Grid: CTH, LTH, KTH, LTU, Tampere, and Aalto. (b) Excerpts of the recordings in a 30 minutes time scale. (c) Zoom into the arrows of panel b of a total length of 2 seconds. (d-f) display the probability density functions (PDF) of the increments Δf_τ at the shortest increment lag $\tau = 0.02$ s and at $\tau = 1$ s, in a vertical logarithmic scale, alongside a normal distribution (which is an inverted parabola in a vertical logarithmic scale) with equal variance for the first recording at $\tau = 0.02$ s, for comparison. PDFs are vertically displaced for clarity. All recordings are synchronous and have a sampling time of 0.02 s.

the bulk frequency without addressing spatial aspects of the dynamics [24–26]. Both spatio-temporal or general time series analysis require access to high-quality, spatially distributed frequency recordings with phasor measurement units (PMUs) or PMU-like devices. Unfortunately, most data sets are not shared openly. While the US-based initiative FNET/GridEye [27] offers updated maps of frequencies world-wide, no open access to the data is available. Other initiatives are not available for the Nordic Grid, such as Grid Radar [28], or only cover spatial measurements in Continental Europe [25, 29].

Within this article, we present self-recorded data of the Nordic synchronous area and power study grid synchronisation in a data-centred model-free approach. We focus on the increment statistics of the frequency time series which carries essential information on the volatility and the synchronisation of the frequency. We show that non-Gaussian increment statistics are ubiquitous in increment statistics and that the variance on the increments scales faster than Brownian-like motions. Next, we show that stochastic fluctuations exhibits spatial correlations between locations even at vanishing time differences and that there exist temporal correlation within the same incremental time series. Finally, we examine phase and amplitude synchronisation separately. To this end, we firstly observe a linear relation in space for phase synchronisation, which contrasts the second finding of a strong (super-)diffusive coupling in amplitude synchronisation.

II. SYNCHRONISATION PHENOMENA AND INCREMENT STATISTICS

In this article we analyse power-grid frequency recordings in the Nordic Grid from the 9th to the 11th of September, 2013, with a sampling time of 0.02 seconds. We illustrate the locations of the recording sites on a map of the Nordic Grid synchronous area in Fig. 1a. An excerpt of 30 minutes of recordings is displayed in b, vertically displaced for clarity. Panel c shows a snapshot of 2 seconds length, around the time point indicated by an arrow in panel b.

One immediately notices the most common properties of power-grid frequency: On course scales, all recordings are synchronous and seem perfectly identical at first sight. If one examines them at a time scale of hundreds of milli seconds (in Fig. 1c) the synchronous behaviour is still present, but small fluctuations occur, each seemingly with their own dynamics, at each location. We will focus on these small, high-frequency fluctuations and introduce here the rather intuitive idea of phase and amplitude synchronisation.

A. Phase and amplitude synchronisation

A coupled dynamical system, say for simplicity, a rotating generator and a (rotating) motor, each with a given inertia, exchanging power, will display some frequency dynamics around the set value. Three different regimes of the collective dynamics exist, as depicted in

Fig. 2 for two sinusoidal signals (e.g. their power-grid frequency). If both signals are only weakly interacting and independent, we would expect these signals to be uncorrelated, i.e., with a different phase and display different amplitudes (panel **a**). Enforcing a synchronised dynamics, e.g. via coupling, will match their phases, but not yet their amplitude. Hence, we observe a correlation between both signals (panel **b**). Lastly, given sufficient coupling, we also obtain amplitude synchronisation, i.e., in this case a full synchronisation, as both phase and amplitude are equal (panel **c**).

Naturally several questions arise from this depiction: What are the usual timescales for each synchronisation to take effect after an external perturbation? What are the causes for this synchronisation? How do distances between elements impact the synchronisation (especially in networked systems)? Lastly, are there properties from one synchronisation phenomenon that affect the other?

We will address these questions on the basis of measured time series. Phase synchronisation is analysed in terms of the signal's increments, i.e., the change of the signal between two points in time. If two signals are perfectly phase synchronised, their increments will perfectly correlated. In contrast, amplitude synchronisation must be studied in terms of the original time series.

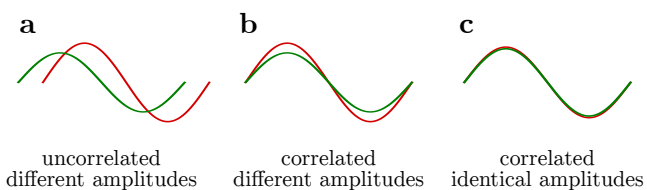


FIG. 2. Distinct frequency signals first become correlated and then converge to identical amplitudes. We illustrate this here: **a**: Initially, signals are initially uncorrelated and have differing amplitudes. **b**: After some time, signals which are phase synchronised but still differ in amplitude. **c**: Finally, signals are synchronised both in phase and amplitude.

B. Increment statistics

In order to study the synchronisation, including correlations of each time series, we investigate the increment statistics of the recordings. Synchronous power-grid systems operate at a set frequency. In the Nordic Grid, and all of Europe, this is the nominal frequency of 50 Hz. Many of the deterministic properties are studied directly from the time series themselves, by either studying their deterministic properties, i.e., as a dynamical system, or as a stochastic process. Yet, here we take another approach and focus on fluctuations of the time series—in particular their increments—and quantify their stochastic properties and correlations.

Increments $\Delta f_\tau(t)$ are defined as

$$\Delta f_\tau(t) = f(t + \tau) - f(t), \quad (1)$$

where the incremental lag τ gives the temporal difference of the two points within a time series $f(t)$. Studying the incremental properties of a time series focuses on the shortest time scales of the underlying processes, thus excludes the deterministic trends and deals solely with the stochastic characteristics of the fluctuations themselves. Note that we move away from considering the recordings in their time domain and study the difference between two points separated by a temporal lag τ .

There is an intrinsic relation between increments and inertia in the system. The ability of a power-grid system to maintain itself at its nominal frequency (50 Hz or 60 Hz) is dictated by the mass of their coupled rotating generators. Short term fluctuations in the power system—i.e., precisely the increments—are conjectured to grow in amplitude and frequency when more inertia is removed as renewable generators replace fossil fuelled ones [4, 5]. These fluctuations are of great importance as they can lead to large angle variations *ergo* power flow fluctuations, putting additional strain on generators and transmission system components.

After obtaining the incremental time series for all six sites, we investigate their statistics for the shortest incremental lag $\tau = 0.02$ s and the longer lag $\tau = 1$ s in Fig. 1d-f. One observes considerable differences of the empirical probability distributions (PDFs) between the six sites. As a base-line we might expect Gaussian distributions, i.e. inverted parabola in the vertical logarithmic plot (dotted line). Indeed, inspecting the incremental distributions at a delay $\tau = 1$ s, some sites, such as CTH, approximately follow such a Gaussian distribution. In contrast, most sites display clear deviations from Gaussianity for the short delay $\tau = 0.02$ s, instead displaying heavy-tailed distributions.

The heavy tails in these distribution indicate that uncommonly large fluctuations take place in the increments, i.e., the difference from one time-point to the next is abnormally large (compared to "normally" distributed noise). Furthermore, we note that these increments seem rather distinct at each location, in particular compared to the much more homogeneous power-grid frequency recordings. This tell us straightforwardly that, although there is a strong synchrony in the system—the 50 Hz of operation—each location varies ever-so-slightly and each in its own manner. What we observe here are local properties at each site of the recordings.

To best quantify both the deviations from Gaussian distributions as well as the differences between sites, we investigate two statistical parameters: the variance and the kurtosis of each distribution, as a function of time step τ .

C. Statistical properties of incremental time series

We examine two statistical moments of the incremental distributions as a function of the incremental lag τ , namely, the variance and kurtosis. The variance (sec-

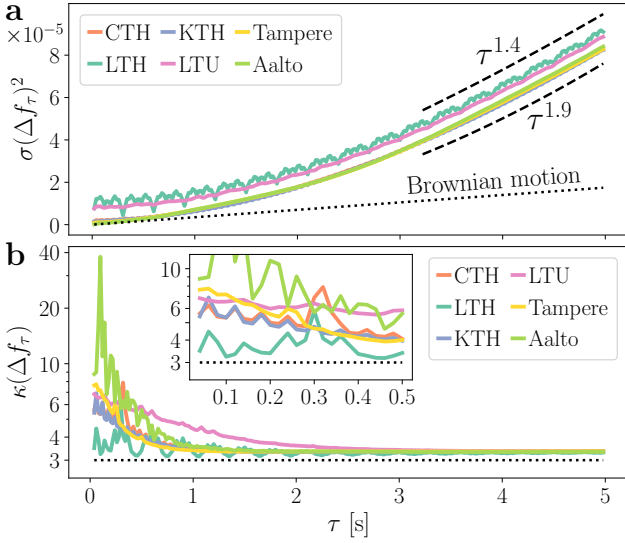


FIG. 3. Variance $\sigma(\Delta f_\tau)^2$ and kurtosis $\kappa(\Delta f_\tau)$ of the increment Δf_τ display clear deviations from Gaussianity. **a** A power-law scaling is observed, with exponent > 1 . For comparison, the linear-like scaling of an uncorrelated Brownian motion is shown. LTU and LTH seem to display the presence of microscopic noise, uniformly increasing the variance of their increments, irrespective of τ . **b** Kurtosis $\kappa(\Delta f_\tau)$ of the increment Δf_τ in a vertical logarithmic scale. The increments statistics are always leptokurtic, i.e., $\kappa > 3$. All increment statistics $\kappa(\Delta f_\tau)$ converge to $\kappa(\Delta f_{\tau \gg 0}) \approx 3.35$ at $\tau \gtrsim 2$.

ond statistical moment) indicates the average displacement of each recording from their mean, i.e., how far the point of the recordings are spread out from their mean. Note that we have a mean zero here, as we investigate frequency fluctuations. On the other hand, the kurtosis (forth normalised moment) roughly indicates how often rather large deviations happen, i.e., deviations much larger than those that fall inside the spread of the standard deviation (square root of the variance). [30].

Note that we handed-picked two distributions at $\tau = 0.02$ s and $\tau = 1$ s for Fig. 1d-g. Now, we systematically investigate how the variance and kurtosis of each increment distribution change as we slowly increase the incremental lag τ . Let us examine variance $\sigma(\Delta f_\tau)^2$ (Fig. 3a) and kurtosis κ Fig. 3b for the first five seconds of incremental lags $\tau \in [0.02 \text{ s}, 5 \text{ s}]$. The first observation is that the variance of the increments Δf_τ increases in a power-like relation. In particular, we observe a scaling of the variance as

$$\sigma(\Delta f_\tau)^2 \sim \tau^{2\alpha}, \text{ with } 1.4 < 2\alpha < 1.9. \quad (2)$$

Notice here that classical Brownian motion, often assumed when simulating noisy processes, e.g. in Matpower or other software, scales with $\sigma(\Delta_{\text{Bm}} f_\tau)^2 \sim \tau$, i.e., ($\alpha = 0.5$), see linear dotted line in Fig. 3a.

This strong diffusion scaling of real noise becomes particularly relevant when performing simulations and

should be incorporated accordingly. Importantly, the observation of this correlated, i.e. non-white noise is ubiquitous and uniform across locations and strictly derived from a data analysis, i.e. without imposing any model. We come back to this observation in Sec. II E and show that this seemingly innocent coefficient plays an decisive role in both phase and amplitude synchronisation.

As a last remark, one observes similarly that in both LTU and LTH additional microscopic noise is present, possibly with some regular properties (notice the curves do not approach zero for decreasing τ). Whether these are artefacts or fundamental physical property of each site's local stochastic properties is left to future analysis.

Let us now turn our attention to how the kurtosis changes with increasing τ (Fig. 3b). As a reference we provide the kurtosis of a Normal distributions as $\kappa_{\mathcal{N}} = 3$ in the plot. Notably, the kurtosis of the increments decreases with growing incremental lags τ , that is to say, as τ increases the distributions become more and more like normal distributions. Phrased differently: On very short time scales, the dynamics displays the highest deviations from Gaussian statistics. One should note that this is not completely unexpected behaviour. In similar analyses on turbulent flows [31–33] and stock market prices [34], similar behaviour is found. What is relevant for our analysis is to understand this as a synchronisation phenomena. As depicted in Fig. 2c, if one is to achieve amplitude synchronisation, then obviously two recordings must end up with identical distributions. This does not mean that the kurtosis of all incremental time series must become normally distributed, but does imply they must be identical.

With these insights into the increment statistics at a single location, we can now address the the first guiding questions formulated in the introduction: Can we unravel the physics of phase synchronisation and its timescale? To do so we must study the smallest timescales of the incremental time series and focus on their correlations.

D. Phase synchronisation

Let us examine the emergence of phase synchronisation of the incremental time series (from uncorrelated to correlated time series). Power grids are designed to ensure that all power generators work synchronously across the power grid. Naturally—and as we have seen so far—this means that the frequency recordings themselves are almost identical, i.e., highly correlated with each other, yet this does not directly translate into fluctuations at each location acting in a similar fashion. Indeed, both models and theory treating the power-grid frequency often assume uncorrelated fluctuations. Again without relying on any model, we will demonstrate that spatial correlations of the fluctuations are ubiquitous, and especially prevalent below a certain distance—similar to a scaling law in long-range fluctuations in power-grid frequency presented in Ref. [35].

In order to study the correlations of the increments be-

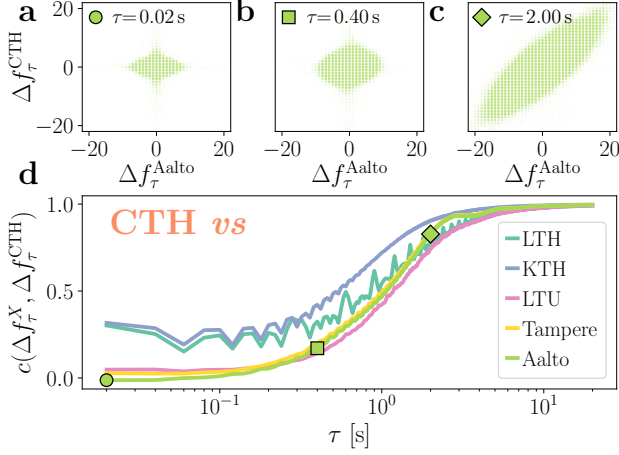


FIG. 4. Pearson correlation $c(\Delta f_\tau^X, \Delta f_\tau^{\text{CTH}})$ of the increments between CTH and the other five locations increases with delay τ . (a-c) display the correlations between CTH and Aalto at **a** $\tau = 0.02$ s, **b** $\tau = 0.40$ s, and **c** $\tau = 2.00$ s. (d) Displays the Pearson correlation $c(\Delta f_\tau^X, \Delta f_\tau^{\text{CTH}})$ in $\tau \in [0.02 \text{ s}, 20 \text{ s}]$. Generally correlations are zero or small with all other recordings at $\tau = 0.02$ s, yet do not vanish for the closest locations to CTH: LTH and KTH. For larger increments, the correlation approaches one: $\lim_{\tau \rightarrow \infty} c = 1$.

tween two locations, we examine the Pearson correlation of two recordings $c(\Delta f_\tau^X, \Delta f_\tau^Y)$

$$c(\Delta f_\tau^X, \Delta f_\tau^Y) = \frac{\text{Cov}(\Delta f_\tau^X, \Delta f_\tau^Y)}{\sigma(\Delta f_\tau^X)\sigma(\Delta f_\tau^Y)} \in [-1, 1], \quad (3)$$

with X and Y the two locations or recordings, σ their individual standard deviation and Cov their covariance. A value of $c = 1$ indicates total correlation, $c = -1$ total anti-correlation, and $c = 0$ no correlation.

Let us take the site CTH as an example. The Pearson correlation $c(\Delta f_\tau^X, \Delta f_\tau^{\text{CTH}})$ with all other locations is displayed in Fig. 4, for $0.02 \text{ s} < \tau < 20 \text{ s}$. Panels **a-c** show how large increments lead to highly aligned, i.e. correlated increments, while short increments display no correlation. The de-correlation of the increments for very small time differences τ is clear, yet this does not seem to be the case for the two closest locations to CTH: LTH and KTH. In this case, the Pearson correlations are $c(\Delta f_\tau^{\text{LTH}}, \Delta f_\tau^{\text{CTH}}) \sim 0.25$ and $c(\Delta f_\tau^{\text{KTH}}, \Delta f_\tau^{\text{CTH}}) \sim 0.25$ at the lower limit of $\tau \rightarrow 0$. All Pearson correlations can be found in App. A, Fig. 7.

This gives rise to the following question: Are fluctuations in the increments correlated for shorter distances, i.e., are there spatial correlations in the increments themselves? We examine this by studying the Pearson correlations against the geographical distance between the locations in two manners: (1) find the Pearson correlation at the lowest temporal lag $\tau = 0.02$ s and the distance between the locations; (2) find how much time it takes for the locations to surpass a given threshold of correlation and the distance between the locations.

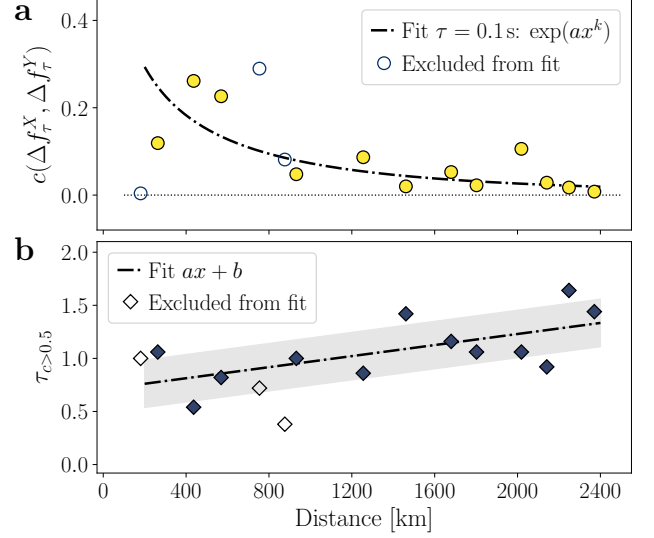


FIG. 5. Time to reach correlation increases linearly with distance. (a) displays the Pearson correlation $c(\Delta f_\tau^X, \Delta f_\tau^Y)$ of two locations at the temporal lag $\tau = 0.1$ s in relation to their driving distance. The presence of non-vanishing correlations at small distances can be observed, following an exponential-like function $\exp(ax^k)$, with parameters $a = -0.10$ and $k = 0.47$. (b) displays the time at which the Pearson correlations between two locations become greater than 0.5, $\tau_{c>0.5}$, plotted against the distance between the locations. In similar fashion, we see that the time it takes for the increments to become correlated is linearly proportional to the distance between two locations. Shade area indicated the standard deviation of the polynomial fitting. Three couplings are discarded: Aalto-Tampere and Aalto-LTU, due to their seemingly small anti-correlation at $\tau = 0.02$, which quickly becomes correlated at $\tau > 0.1$ s, and the LTH-KTH coupling, which as seen in Fig. 3a present microscopic noise, mooting possible correlations.

Firstly we note that all sea cables in the Baltic Sea or the Gulf of Bothnia are direct current (DC) cables, thus these do not participate in the synchrony on the system. That being the case, driving/walking distances between two locations serves as a proxy to actual grid distance, i.e., the actual power-line cable lengths between two locations.

In Fig. 5a, the fifteen distance pairs between the six locations and their respective Pearson correlation at temporal lag $\tau = 0.1$ s are displayed. The Pearson correlations $c(\Delta f_\tau^X, \Delta f_\tau^Y)$ for all locations at $\tau = 0.1$ s are plotted against the distances between the points. A pattern of correlations emerges for short distances with non-vanishing Pearson correlations. Three couplings are discarded: (1) Aalto-Tampere and Aalto-LTU, due to their small anti-correlation at very small $\tau = 0.02$ s–0.1 s. (2) the LTU-LTH coupling, where the present microscopic noise moots evaluating correlations (see e.g. Fig. 3a).

In Fig. 5b we examine the speed at which correlations between two locations become larger than 0.5 (50%), i.e., the find the time $\tau_{c>0.5}$ at which the Pearson correla-

tions $c(\Delta f_\tau^X, \Delta f_\tau^Y) > 0.5$, in relation to the distance between the locations. We observe that locations closer to each other see their increments becoming correlated faster than locations that are far apart. While this general trend is expected, note the linear relation between phase-synchronisation and geographical distance between the sites. We will have cause to contrast this with amplitude synchronisation in the following section. Furthermore, notice that this phenomena is strictly bound to taking place < 2 s, i.e., phase synchronisation takes place at very short temporal scales.

We now move forth to examine the emergence of amplitude synchronisation.

E. Amplitude synchronisation

We have seen that phases of increments synchronised within < 2 s, so let us now move to amplitude synchronisation. As we have seen in Fig. 3a the variance of the increment statistics increases in a power-law relation to the incremental lag τ . In order to best describe this—and to retrieve the scaling constant α —we employ a Detrended Fluctuation Analysis (DFA) of the power-grid frequency recordings [36–38]. Recall that DFA studies the scaling of the fluctuations of a time series by studying the local properties of the data, in a similar fashion to what increment statistics does. With DFA we extract the fluctuation function $F^2(r)$ over a scale r of a time series, in a much similar fashion to our incremental lag τ . We will keep τ and r distinct since the DFA is applied directly to the power-grid frequency recordings, not the incremental time series. Still, DFA and increment analysis are intrinsically related and studying the fluctuation function $F^2(r)$ in relation to the scale r will help us uncover the scaling parameter α , as also observed in Fig. 3a.

The DFA procedure—applied on the power-grid frequency recordings—is the following: Take non-overlapping segments of the power-grid frequency, fit a polynomial function (of order one, in this case), subtract the fit from the segment of data, and extract the variance (i.e., fluctuation function $F^2(r)$) of each detrended segment. By increasing the segment size one can study the change of the variances as a function of the segment size (the scale r). If the time series follows a power-law, which we show in Fig. 3a it does, one can evaluate the plots of the fluctuation function $F^2(r)$ in the scale r in a double logarithmic scale. The reason to compute the fluctuation function $F^2(r)$ instead of examining the original time series is simple: Actual power-grid frequency recording have trends, such as short-terms jumps due to dispatch and market activity or a permanent small mismatch of power generation and consumption. We wish to disentangle these trends from the true underlying fluctuation dynamics, which DFA is capable of.

For our purpose here, we have already established that the incremental time series follow a diffusion scale with a power-law like distribution given by Eq. (2), which is

manifestly larger in exponent than a Brownian motion. We plot the fluctuation function $F^2(r)$ of the six time series in a scale $r \in [1 \text{ s}, 20 \text{ s}]$ in Fig. 6a, normalised by the scaling of a Brownian motion, which has a scaling power $\alpha^{\text{Bm}} = 1.5$. We fit the difference in scaling, represented by $\alpha' = \alpha - \alpha^{\text{Bm}}$. We find that $\alpha' \approx 0.376$, i.e., $\alpha = 1.876$, for the range of $r \in [5 \text{ s}, 10 \text{ s}]$. Noticeable is also the distinct separation of the curves at timescales $r < 5$ s, which has been identified in Ref. [35] as a possible method to study the synchronisation of fluctuation over spatial distances. We compute the relative DFA function $\eta(r)$, as introduced in Ref. [35], with CTH as a reference in Fig. 6b. This function represents the relative variations of the fluctuations to a reference point, given by

$$\eta(r)^Y = \frac{F^2(r)^X - F^2(r)^Y}{F^2(r)^Y}, \quad (4)$$

with Y the reference location, CTH in our case, and X the remaining locations. The synchronous nature of power-grids ensures all fluctuations eventually collapse into a single function, i.e., all amplitudes become identical. This also implies that, in general, a set of recordings with a temporal scale > 5 s anywhere in the Nordic grid are indistinguishable.

To quantify this synchronisation, we define a ‘time-to-bulk’ χ as follows. We record the time it takes each time series’ relative fluctuations η to be reduced to 0.1 (10%) and thereby become indiscernible from the reference (indicated by the grey line in Fig 6b). Using CTH as a reference, we obtain χ^{CTH} and observe how the time towards synchronisation increases with distance in Fig. 6c. An intuitive assumption is to compare the empirical data with diffusive-like behaviour, i.e., a fit of quadratic order. While this captures the general trend, it provides a sub-optimal fit. Alternatively, we consider a (super-)diffusion fit where we propose the ‘time-to-bulk’ χ scales with the distance to the power of 2α . Indeed, this seems more adequate to describe the data, suggesting the presence of correlations in the noise structure (given by α) influence the speed at which synchronisation of the fluctuations is achieved in space.

We pause here and must thoroughly examine what the ‘time-to-bulk’ analysis just uncovered. First, we confirmed the suspicion raised in Fig. 3a that the power-grid frequency indeed exhibits temporally correlated noise, i.e., the assumption that the high-frequency fluctuations in power-grid frequency recordings are purely white noise is not justified. In fact, $\alpha > 1.5$ indicates the present of positively correlated motions, which show a power-law like diffusion, as seen in Fig. 3a. One finds that all time series coalesce to a Hurst index of $H = \alpha - 1 = 0.876$, i.e., strongly positively correlated noise. This Hurst index can be determined efficiently for any given power-grid frequency measurement [38] and should be incorporated in stochastic studies of power systems.

Secondly, we observe that these fluctuations display a spatial relation between the locations, i.e., locations which are closer see their fluctuation amplitude synchro-

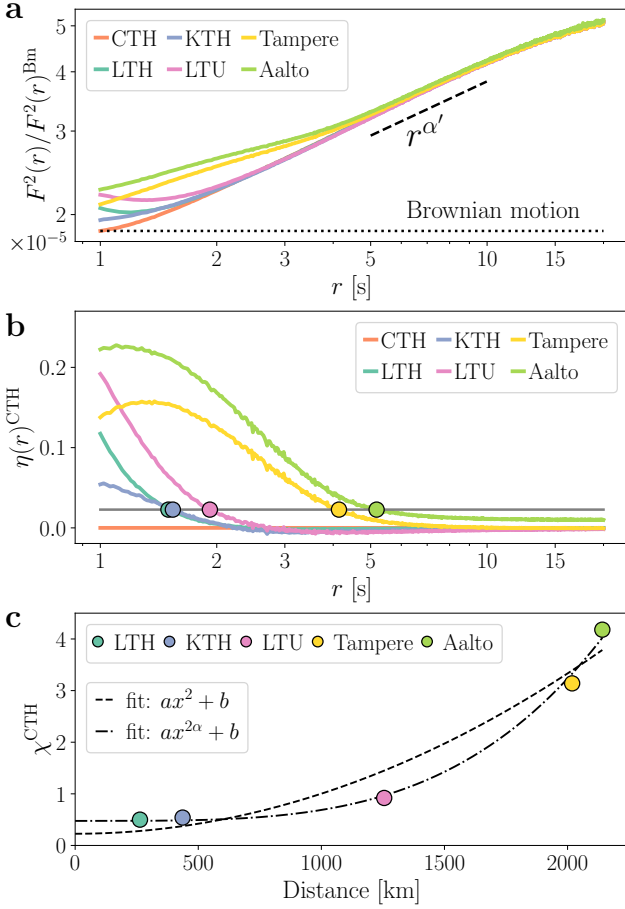


FIG. 6. Correlated, diffusive coupling revealed by DFA, relative DFA $\eta(r)$, and ‘time-to-bulk’ χ . (a) displays the normalised fluctuation function $F^2(r)$ obtained from DFA with first-order polynomials, normalised by the fluctuation function $F^2(r)^{Bm}$ of a Brownian motion ($\alpha - 1 = 0.5$). The scaling $\alpha' = 0.376$, leading to an over scaling $\alpha = 1.876$, or equivalently a Hurst index $H = \alpha - 1 = 0.876$, i.e., strongly positively correlated noise. (b) displays the relative fluctuation function $\eta(r)^{CTH}$, with CTH as reference. The fluctuations of the other five locations decay in time. When indicated markers are the points (in time) where the relative size of the fluctuations hits 0.1 (10%) of the maximal relative fluctuation. (c) shows the ‘time-to-bulk’ function χ^{CTH} against the distances of each recordings’ location in relation to CTH. A diffusive-like behaviour, i.e., a fit of quadratic order, with the distance is compared with a super-diffusive fit of order 2α .

nise faster than those farther apart. Note that this a second effect, different from the (temporal) correlation in the time series, which we have examined before. Focusing on the new spatial correlations, we know all power-grid frequency recordings are highly correlated, they practically follow the same phenomena at timescales > 5 s, i.e., their dynamical oscillations around 50 Hz. Moreover, we have seen above that the correlations of the increments are already large at these long timescales, while displaying zero or low correlations on short time scales. What

we now observe is a synchronisation of the amplitudes of the fluctuations as a function of time. If we now compare Fig. 6 to Fig. 5, we notice a sharp difference in the physics of the synchronisation phenomena: The amplitude synchronisation (Fig. 6) is achieved with a large power-like relation with exponent 2α , much in contrast with the linear relation for phase synchronisation (Fig. 5).

Yet, one more property is of paramount relevance. We observe an amplitude synchronisation as a function of the distance to the power 2α , intrinsically linking local properties with the overall synchronisation phenomena in space. Notice that we initially uncovered the scaling exponent α from the variances of the incremental time series. This α is a uniquely temporal property of the noise of a given time series—with no explicit spatial connection between locations. Using DFA, we revealed that this seemingly local exponent α plays a crucial role on the rate at which spatial amplitude synchronisation takes place. Take, for example, that α would be 1.5 ($H = \alpha - 1 = 0.5$), i.e., we would have a truly Brownian motion. By this relation, a conceptual location at 3000 km from CTH should take ~ 7.2 s to achieve amplitude synchronisation. If now we take the empirical $\alpha = 1.876$, this amplitude synchronisation is only achieved at ~ 13 s. The delay of global amplitude synchronisation by local correlations becomes even more noticeable for longer distances.

We note that *a priori* there is no necessary relation between local temporal properties of the incremental time series and the spatial convergence of all oscillations into one bulk behaviour. Nevertheless, that local temporal properties affect the spatial correlations of systems that rely heavily in synchronisation to operate is legitimate, yet has not been described before.

III. CONCLUSION

In this article we have analysed synchronous recordings of the power-grid frequency from six locations in the Nordic Grid from September 2013. The high temporal resolution of these recordings of just 0.02 s allows for a detailed analysis of power system synchronisation via the statistics and correlations of the increments. Essential insights into the temporal and spatial scales of synchronisations can be extracted from the ambient fluctuations in a model-free approach. Our results further emphasise the outstanding importance of a broad availability of high-quality data for research on power system operation and energy science in general [39].

We investigated the distribution of increments (frequency differences) and noted severe deviations from Gaussianity. In particular, increment distributions are highly leptokurtic, i.e. display heavy tails. Noticeable differences in the increment statistics at each six locations are observed, indicating that the incremental time series reflect above all the local phenomena of power generation and consumption in the location of the recording. We saw as well a relaxation of the incremental time se-

ries kurtosis for large time lags τ , yet even for very large lags we observe a leptokurtic distribution with kurtosis of $\kappa(\Delta f_{\tau \gg 0}) \approx 3.35$.

The emergence of phase synchronisation is revealed by the correlations of the increments on different temporal and spatial scales. Our analysis has shown two essential results: (1) If locations are rather close, the increments are correlated even at the shortest possible time lag of 0.02 s, indicating that ambient fluctuations of power generation and consumption that drive the frequency dynamics are correlated. In light of this result, any assumption of spatially independent noise in power system simulations should be carefully reviewed. (2) The increments at locations get become strongly correlated on a time scale of one second depending on the distance of the location. The further two locations are from each other, the longer it takes before correlations exceed a certain value.

Strong temporal correlations in the fluctuations are revealed by the variance of the incremental time series reveals. We find that the variance follows a power-law $\sim \tau^{2\alpha}$ as a function of the incremental lag τ , with an exponent much higher than for ordinary uncorrelated Brownian motion. This result is confirmed by a detrended fluctuation analysis (DFA), which yields the exponent $2\alpha = 1.876$, i.e., a positively correlated Hurst index $H = 0.876$. DFA is further used to study the time scales of amplitude synchronisation by quantifying the time needed until the fluctuation functions at two locations become similar up to a certain level. This ‘time-to-bulk’ function scales as a power law with the distance with an exponent 2α larger than one, hence faster than in an ordinary diffusion process. This is particularly relevant as it couples a temporal property of the incremental time series, i.e., the scaling of the variance, with a spatial property of the amplitude of the fluctuations, suggesting that local temporal correlations enforces a global scale for spatial synchronisation. A diffusive scaling was initially proposed for amplitude synchronisation in Ref. [35], yet here we argue further that the local temporal properties might influence the global synchronisation phenomena even faster than a regular diffusion. Interestingly, the linear scaling of phase synchronisation, compared to the power law scaling of amplitude synchronisation implies that locations far apart will first synchronise in their phase, then amplitude, while the ordering might be reversed for locations geographically nearby. A definite answer will require data from more locations, preferably in large synchronous grids.

We remark here that the implications of these findings are considerable for both the understanding of the physical phenomena behind power-grid systems and for simulations. We conclude that the implications of these findings are considerable for both the understanding of the dynamical processes in power grids as well as their simulation, as the increments provide as a proxy for fluctuations of the power imbalance that drives the frequency dynamics. Foremost, we uncovered that the temporal stochastic properties at each recording site impacts the

speed at which amplitude synchronisation is achieved. This grants an exact measure for amplitude synchronisation across any power-grid, which can now be uncovered solely from one single local recording, i.e., since the scaling parameter α is a local property, a single power-grid frequency measurement in a synchronous region allows us to determine at which speed a far-away location will be amplitude-synchronised with the rest of the grid. Furthermore, two distinct and so far practically unaddressed characteristics are uncovered: First, fluctuations at different locations are correlated. This implies immediately that any simulation—especially for small power grids as microgrids—must consider the presence not only of noise, i.e., stochastic fluctuations, but of spatially correlated noise. We observe that substantial correlations are seen up to distances of 1000 km. Secondly, each location, each power-grid frequency recordings, shows distinct strong temporal correlation within itself. This is a clear indication that temporal correlations in the stochastic fluctuations are present. In fact, these correlations dictate the physics of the aforementioned amplitude synchronisation. Thus, adequate simulations should extent their analysis far beyond classical white noise (uncorrelated Brownian noise) and consider instead spatio-temporally correlated (non-white) noise.

All of these effects significantly impact risk assessment: Heavy-tailed noise leads to larger deviations than expected from Gaussian noise, while correlated noise continues to push the system in one given direction, instead of randomly fluctuating around zero. Hence, both correlations and non-Gaussian distributions induce larger frequency deviations than assumed from white Gaussian noise and thereby increase the risk of destabilising the system. Simply adding a larger security margin when carrying out simulations is barely appropriate: A broad Gaussian noise distribution would imply that medium deviations occur very often when instead rare large deviations are the problem. Hence, to properly dimension back-up capacities and design adequate control options, the non-standard statistics presented here should be included.

In the future, it would be desirable to study the link between local temporal correlation and spatial amplitude synchronisation in different data sets. Also, further theoretical work is necessary to offer efficient simulation tools including all discussed phenomena.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the German Federal Ministry of Education and Research (grant no. 03EK3055B) and the Helmholtz Association (via the joint initiative “Energy System 2050 – A Contribution of the Research Field Energy” and the grant “Uncertainty Quantification – From Data to Reliable Knowledge (UQ)” with grant no. ZT-I-0029). This work was performed as part of the Helmholtz School for Data Sci-

ence in Life, Earth and Energy (HDS-LEE). This project has received funding from the European Union's Hori-

zon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840825.

-
- [1] J. Machowski, Z. Lubosny, J. W. Bialek, and J. R. Bumby, *Power System Dynamics: Stability and Control*, 3rd ed. (John Wiley & Sons, 2020).
 - [2] M. Klein, G. J. Rogers, and P. Kundur, A fundamental study of inter-area oscillations in power systems, *IEEE Transactions on power systems* **6**, 914 (1991).
 - [3] UCTE, Final report system disturbance on 4 november 2006, https://www.entsoe.eu/fileadmin/user_upload/_library/publications/ce/otherreports/Final-Report-20070130.pdf (2007).
 - [4] A. Ulbig, T. S. Borsche, and G. Andersson, Impact of low rotational inertia on power system stability and operation, *IFAC Proceedings Volumes* **47**, 7290 (2014).
 - [5] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbič, Foundations and challenges of low-inertia systems, in *2018 Power Systems Computation Conference (PSCC)* (IEEE, 2018) pp. 1–25.
 - [6] F. Dörfler and F. Bullo, Synchronization in complex networks of phase oscillators: A survey, *Automatica* **50**, 1539 (2014).
 - [7] F. Dörfler and F. Bullo, Synchronization and transient stability in power networks and nonuniform kuramoto oscillators, *SIAM Journal on Control and Optimization* **50**, 1616 (2012).
 - [8] F. Dörfler, M. Chertkov, and F. Bullo, Synchronization in complex oscillator networks and smart grids, *Proceedings of the National Academy of Sciences* **110**, 2005 (2013).
 - [9] J. Schiffer, R. Ortega, A. Astolfi, J. Raisch, and T. Sezi, Conditions for stability of droop-controlled inverter-based microgrids, *Automatica* **50**, 2457 (2014).
 - [10] M. Colombino, D. Groß, J.-S. Brouillon, and F. Dörfler, Global phase and magnitude synchronization of coupled oscillators with application to the control of grid-forming power inverters, *IEEE Transactions on Automatic Control* **64**, 4496 (2019).
 - [11] H.-D. Chiang, F. Wu, and P. Varaiya, Foundations of direct methods for power system transient stability analysis, *IEEE Transactions on Circuits and systems* **34**, 160 (1987).
 - [12] M. D. Omar Faruque, T. Strasser, G. Lauss, V. Jalili-Marandi, P. Forsyth, C. Dufour, V. Dinavahi, A. Monti, P. Kotsampopoulos, J. A. Martinez, K. Strunz, M. Saeedifard, Xiaoyu Wang, D. Shearer, and M. Paolone, Real-time simulation technologies for power systems design, testing, and analysis, *IEEE Power and Energy Technology Systems Journal* **2**, 63 (2015).
 - [13] B. Lu, X. Wu, H. Figueroa, and A. Monti, A low-cost real-time hardware-in-the-loop testing approach of power electronics controls, *IEEE Transactions on Industrial Electronics* **54**, 919 (2007).
 - [14] G. F. Lauss, M. O. Faruque, K. Schoder, C. Dufour, A. Viehweider, and J. Langston, Characteristics and design of power hardware-in-the-loop simulations for electrical power systems, *IEEE Transactions on Industrial Electronics* **63**, 406 (2015).
 - [15] A. R. Messina, *Inter-area Oscillations in Power Systems: A Nonlinear and Nonstationary Perspective*, 1st ed. (Springer US, 2009).
 - [16] M. Feldman, Hilbert transform in vibration analysis, *Mechanical Systems and Signal Processing* **25**, 735 (2011).
 - [17] J. L. Rueda, C. A. Juarez, and I. Erlich, Wavelet-based analysis of power system low-frequency electromechanical oscillations, *IEEE Transactions on Power Systems* **26**, 1733 (2011).
 - [18] K. Uhlen, S. Elenius, I. Norheim, J. Jyrinsalo, J. Eloväärä, and E. Lakervi, Application of linear analysis for stability improvements in the nordic power transmission system, in *2003 IEEE Power Engineering Society General Meeting*, Vol. 4 (2003) pp. 2097–2103.
 - [19] L. Vanfretti, L. Dosiek, J. W. Pierre, D. Trudnowski, J. H. Chow, R. García-Valle, and U. Aliyu, Application of ambient analysis techniques for the estimation of electromechanical oscillations from measured pmu data in four different power systems, *European Transactions on Electrical Power* **21**, 1640 (2011).
 - [20] L. Vanfretti, R. García-Valle, K. Uhlen, E. Johansson, D. Trudnowski, J. W. Pierre, J. H. Chow, O. Samuelsson, J. Østergaard, and K. E. Martin, Estimation of eastern denmark's electromechanical modes from ambient phasor measurement data, in *2010 IEEE Power Engineering Society General Meeting* (2010) pp. 1–8.
 - [21] K. Uhlen, L. Vanfretti, M. M. de Oliveira, A. B. Leirbukt, V. H. Aarstrand, and J. O. Gjerde, Wide-area power oscillation damper implementation and testing in the norwegian transmission network, in *2012 IEEE Power and Energy Society General Meeting* (2012) pp. 1–7.
 - [22] J. W. Pierre, D. J. Trudnowski, and M. K. Donnelly, Initial results in electromechanical mode identification from ambient data, *IEEE Transactions on Power Systems* **12**, 1245 (1997).
 - [23] G. Papaefthymiou, P. Schavemaker, L. van der Sluis, W. Kling, D. Kurowicka, and R. Cooke, Integration of stochastic generation in power systems, *International Journal of Electrical Power & Energy Systems* **28**, 655 (2006).
 - [24] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme, Non-gaussian power grid frequency fluctuations characterized by lévy-stable laws and superstatistics, *Nature Energy* **3**, 119 (2018).
 - [25] L. Rydin Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, and B. Schäfer, Data-driven model of the power-grid frequency dynamics, *IEEE Access* **8**, 43082 (2020).
 - [26] M. Anvari, L. R. Gorjão, M. Timme, D. Witthaut, B. Schäfer, and H. Kantz, Stochastic properties of the frequency dynamics in real and synthetic power grids, *Physical Review Research* **2**, 013339 (2020).
 - [27] Power Information Technology Lab, University of Tennessee, Knoxville and Oak Ridge National Laboratory, FNET/GridEye, <http://fnetpublic.utk.edu/> (2014).
 - [28] MagnaGen GmbH, Gridradar—An Independent Grid Monitoring System, <https://gridradar.net/> (2020).
 - [29] R. Jumar, H. Maass, B. Schäfer, L. R. Gorjão, and V. Hagenmeyer, Power grid frequency data base, arXiv

- preprint arXiv:2006.01771 (2020).
- [30] P. H. Westfall, Kurtosis as peakedness, 1905–2014. RIP, *The American Statistician* **68**, 191 (2014).
 - [31] B. Castaing, Y. Gagne, and E. Hopfinger, Velocity probability density functions of high Reynolds number turbulence, *Physica D: Nonlinear Phenomena* **46**, 177 (1990).
 - [32] B. Castaing, Scalar intermittency in the variational theory of turbulence, *Physica D: Nonlinear Phenomena* **73**, 31 (1994).
 - [33] C. Beck and E. G. D. Cohen, Superstatistics, *Physica A* **322**, 267 (2003).
 - [34] D. Xu and C. Beck, Transition from lognormal to χ^2 -superstatistics for financial time series, *Physica A: Statistical Mechanics and its Applications* **453**, 173 (2016).
 - [35] L. R. Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, G. C. Yalcin, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer, Open data base analysis of scaling and spatio-temporal properties of power grid frequencies, *Nature Communication* **11**, 6362 (2020).
 - [36] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Mosaic organization of DNA nucleotides, *Physical Review E* **49**, 1685 (1994).
 - [37] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series, *Physica A* **316**, 87 (2002).
 - [38] L. Rydin Gorjão, MFDFA: Multifractal Detrended Fluctuation Analysis in Python, <https://zenodo.org/record/3625759> (2020).
 - [39] S. Pfenninger, Energy scientists must show their workings, *Nature* **542**, 393 (2017).

Appendix A: Pearson correlation of increments of power-grid frequency recordings

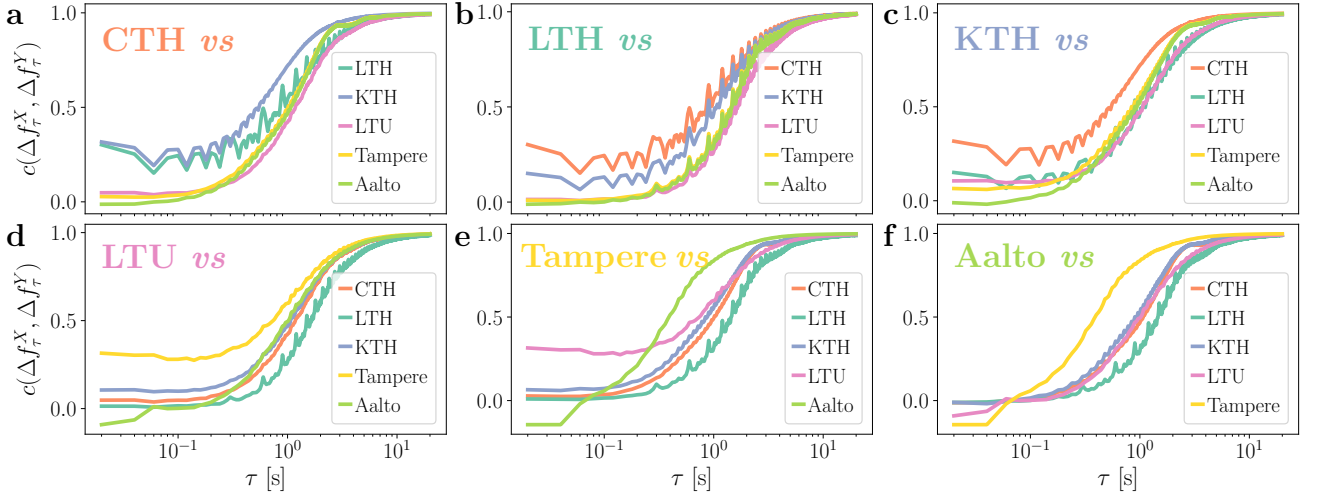


FIG. 7. Pearson correlation $c(\Delta f_\tau^X, \Delta f_\tau^Y)$ of increments of power-grid frequency recordings in for the six locations, for $0.02 \text{ s} < \tau < 20 \text{ s}$.

2.3 Jump-diffusion processes and analysis of paleo-climatic transitions and Dansgaard–Oeschger events

2.3.1 Publication #6

L. Rydin Gorjão, J. Heysel, K. Lehnertz, and M. R. R. Tabar. *Analysis and data-driven reconstruction of bivariate jump-diffusion processes*. Physical Review E **100**, 2019, p. 062127, Ref. [6].

Status: published

Analysis and data-driven reconstruction of bivariate jump-diffusion processesLeonardo Rydin Gorjão^{1,2,3,4,*} Jan Heysel^{1,2,†} Klaus Lehnertz^{1,2,5,‡} and M. Reza Rahimi Tabar^{6,7,§}¹*Department of Epileptology, University of Bonn, Venusberg Campus 1, 53127 Bonn, Germany*²*Helmholtz Institute for Radiation and Nuclear Physics, University of Bonn, Nussallee 14-16, 53115 Bonn, Germany*³*Forschungszentrum Jülich, Institute for Energy and Climate Research—Systems Analysis and Technology Evaluation (IEK-STE), 52428 Jülich, Germany*⁴*Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany*⁵*Interdisciplinary Centre for Complex Systems, University of Bonn, Brühler Straße 7, 53175 Bonn, Germany*⁶*Institute of Physics and ForWind, Carl von Ossietzky University of Oldenburg, Carl-von-Ossietzky-Straße 9-11, 26111 Oldenburg, Germany*⁷*Department of Physics, Sharif University of Technology, 11365-9161 Tehran, Iran*

(Received 27 September 2019; published 20 December 2019)

We introduce the bivariate jump-diffusion process, consisting of two-dimensional diffusion and two-dimensional jumps, that can be coupled to one another. We present a data-driven, nonparametric estimation procedure of higher-order (up to 8) Kramers-Moyal coefficients that allows one to reconstruct relevant aspects of the underlying jump-diffusion processes and to recover the underlying parameters. The procedure is validated with numerically integrated data using synthetic bivariate time series from continuous and discontinuous processes. We further evaluate the possibility of estimating the parameters of the jump-diffusion model via data-driven analyses of the higher-order Kramers-Moyal coefficients, and the limitations arising from the scarcity of points in the data or disproportionate parameters in the system.

DOI: [10.1103/PhysRevE.100.062127](https://doi.org/10.1103/PhysRevE.100.062127)**I. INTRODUCTION**

Research over the last two decades has demonstrated the high suitability of the network paradigm in advancing our understanding of natural and man-made complex dynamical systems [1–7]. With this paradigm, a system component is represented by a vertex and interactions between components are conveyed by edges connecting vertices, and graph theory provides a large repertoire of methods to characterize networks on various scales.

Characterizing properties of interactions using the knowledge of the dynamics of each of the components is key to understanding real-world systems. To achieve this goal, a large number of time-series-analysis methods have been developed that originate from synchronization theory, nonlinear dynamics, information theory, and statistical physics (for an overview, see Refs. [8–15]). Some of these methods make rather strict assumptions about the dynamics of network components generating the time series and many approaches preferentially focus on the low-dimensional deterministic part of the dynamics.

Real-world systems, however, are typically influenced by random forcing, and interactions between constituents are highly nonlinear, which results in very complex, stochastic, and nonstationary system behavior that exhibits both deterministic and stochastic features. Aiming at determining

characteristics and strength of fluctuating forces as well as at assessing properties of nonlinear interactions, the analysis of such systems is associated with the problem of retrieving a stochastic dynamical system from measured time series. There is a substantial existing literature [16–19] for the modeling of complex dynamical systems which employs the conventional Langevin equation that is based on the first- and second-order Kramers-Moyal (KM) coefficients, known as drift and diffusion terms. All functions and parameters of this modeling can be found directly from the measured time series employing a widely used nonparametric approach. There are by now only few studies that make use of this ansatz to characterize interactions between stochastic processes [20–24].

Despite its successful application in diverse scientific fields, growing evidence indicates that the continuous stochastic modeling of time series of complex systems (the white-noise-driven Langevin equation) should account for the presence of discontinuous jump components [19,25–33]. In this context, the jump-diffusion model [34–37] was shown to provide a theoretical tool to study processes of known and unknown nature that exhibit jumps. It allows one to separate the deterministic drift term as well as different stochastic behaviors, namely, diffusive and jumpy behavior [19,32,33]. Moreover, all of the unknown functions and coefficients of a dynamical stochastic equation that describe a jump-diffusion process can be derived directly from measured time series. This approach involves estimating higher-order (≥ 3) KM coefficients and it provides an intuitive physical meaning of these coefficients.

The focus of this paper is to introduce a method to investigate bivariate time series with discontinuous jump

*l.rydin.gorjao@fz-juelich.de

†jan.heysel@uni-bonn.de

‡klaus.lehnertz@ukbonn.de

§tabar@uni-oldenburg.de

components. We begin with an overview of bivariate diffusion processes that exhibit the known relation between the parameters and functions in stochastic modeling and the KM coefficients. Exemplary processes are portrayed, and we propose a measure to judge the quality of our reconstruction procedure. We then present bivariate jump-diffusion processes alongside the associated KM expansion [32], and investigate the suitability of our reconstruction procedure using various examples. We conclude this paper by summarizing our findings.

II. BIVARIATE JUMP-DIFFUSION MODEL

A bivariate jump-diffusion process can be modeled via [19,32]

$$\begin{aligned} \overbrace{\begin{pmatrix} dy_1(t) \\ dy_2(t) \end{pmatrix}}^y &= \underbrace{\begin{pmatrix} N_1 \\ N_2 \end{pmatrix}}_{\text{drift}} dt + \underbrace{\begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix}}_{\text{diffusion}} \underbrace{\begin{pmatrix} dw_1 \\ dw_2 \end{pmatrix}}_{d\mathbf{w}} \\ &+ \underbrace{\begin{pmatrix} \xi_{1,1} & \xi_{1,2} \\ \xi_{2,1} & \xi_{2,2} \end{pmatrix}}_{\text{Poissonian jumps}} \underbrace{\begin{pmatrix} dJ_1 \\ dJ_2 \end{pmatrix}}_{d\mathbf{J}}, \end{aligned} \quad (1)$$

where all the elements of vectors \mathbf{N} , $d\mathbf{J}$, and $d\mathbf{w}$ as well as of matrices \mathbf{g} and ξ may, in general, be state and time dependent (dependencies not shown for convenience of notation). The drift coefficient is a two-dimensional vector $\mathbf{N} = (N_1, N_2)$ with $\mathbf{N} \in \mathbb{R}^2$, where each dimension of \mathbf{N} , i.e., N_i , may depend on $y_1(t)$ and $y_2(t)$. The diffusion coefficient takes a matrix $\mathbf{g} \in \mathbb{R}^{2 \times 2}$. The two Wiener processes $\mathbf{w} = (w_1, w_2)$ act as independent Brownian noises for the state variables $y_1(t)$ and $y_2(t)$. The diagonal elements of \mathbf{g} comprise the diffusion coefficients of self-contained stochastic diffusive processes, and the off-diagonal elements represent interdependencies between the two Wiener processes, i.e., they result from an interaction between the two processes. Each single-dimensional stochastic process element dw_i of $d\mathbf{w}$ is an increment of a Wiener process, with $\langle dw_i \rangle = 0$, $\langle dw_i^2 \rangle = dt$, $\forall i$. The discontinuous jump terms are contained in $\xi \in \mathbb{R}^{2 \times 2}$ and $d\mathbf{J} \in \mathbb{N}^2$, where $d\mathbf{J}$ represents a two-dimensional Poisson process. These are Poisson-distributed jumps with an average jump rate $\lambda \in \mathbb{R}^2$ in unit time t . The average expected number of jumps of each jump process J_i in a timespan t is $\lambda_i t$. The jump amplitudes ξ are Gaussian distributed with zero mean and standard deviation $\xi_{i,j}$.

We here consider merely autonomous systems. Nonergodic problems are beyond the scope of this paper, and a more delicate approach to both bivariate stochastic processes would be needed.

III. BIVARIATE DIFFUSION PROCESSES

Let us begin with bivariate diffusion processes, for which the model takes the form

$$\overbrace{\begin{pmatrix} dy_1(t) \\ dy_2(t) \end{pmatrix}}^y = \underbrace{\begin{pmatrix} N_1 \\ N_2 \end{pmatrix}}_{\text{drift}} dt + \underbrace{\begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix}}_{\text{diffusion}} \underbrace{\begin{pmatrix} dw_1 \\ dw_2 \end{pmatrix}}_{d\mathbf{w}}. \quad (2)$$

The model consists of six functions, two for the drift coefficients and four for the diffusion coefficients. Given a bivariate diffusion process, can we reconstruct the aforementioned parameters strictly from data? Extensive work exists on this matter [18], especially covering purely diffusion processes, and we will use these now as a stepping stone to jump-diffusion processes. Understanding the working and contingencies of reconstructing the parameters of a diffusion process [Eq. (2)] will serve as a gateway to understand how a similar procedure awards equivalent measures for jump-diffusion processes. We address the aforementioned question first by revisiting the mathematical foundation that allows one to recover, strictly from data, the drift \mathbf{N} and diffusion \mathbf{g} coefficients. Subsequently, we numerically integrate diffusion processes with *a priori* fixed values of the drift \mathbf{N} and diffusion \mathbf{g} coefficients and aim at retrieving these values strictly from the generated data (the Euler-Mayurama scheme with a time sampling of 10^{-3} over a total of 10^5 time units, i.e., 10^8 number of data points). If the actual and retrieved values match, the reconstruction method is effective.

A stochastic process has a probabilistic description given by the master equation [16,19]. It does not describe a specific stochastic process in itself, but the probabilistic evolution of the process in time. The master equation accepts an expansion in terms, the KM expansion, that allows for a purely differential description of the process. More importantly, the coefficients of the expansion, known as KM coefficients, entail directly a relation to the aforementioned parameters of a stochastic process given by Eq. (1). The exact relation will be given below. There is, though, an important detail regarding the KM coefficients: they are in themselves not constants but functions on the underlying space or, in other words, a scalar field, and for our purposes here they can be understood as two-dimensional surfaces. We will denote these as KM surfaces.

Lastly, and more familiar, the Fokker-Planck equation is a truncation of the KM expansion at second order. It is especially relevant given its connection to physical processes and the Pawula theorem [38]. The Pawula theorem ensures that the truncation is not ill suited for the underlying process if the fourth-order KM coefficient approaches zero in the limit $dt \rightarrow 0$. It is now crucial to understand that the theorem holds for a one-dimensional process, and we are not aware of a proof for higher dimensions. This contrasts the common notion that studying only the first two KM coefficients of two- or higher-dimensional processes is sufficient (see Refs. [32,33,39] and references therein).

The KM coefficients $\mathcal{M}^{[\ell,m]}(x_1, x_2) \in \mathbb{R}^2$ of orders (ℓ, m) are defined as

$$\begin{aligned} \mathcal{M}^{[\ell,m]}(x_1, x_2) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int [y_1(t + \Delta t) - y_1(t)]^\ell [y_2(t + \Delta t) - y_2(t)]^m \\ &\quad \times P(y_1, y_2; t + \Delta t | y_1, y_2; t) |_{y_1(t)=x_1, y_2(t)=x_2} dy_1 dy_2 \end{aligned}$$

and can be obtained from bivariate time series $(y_1(t), y_2(t))$. Theoretically, Δt should take the limiting case of $\Delta t \rightarrow 0$, but the restriction of any measuring or storing devices—or the nature of the observables themselves—permits only time-sampled or discrete recordings. The relevance and importance

of adequate time sampling was extensively studied and discussed in Refs. [19,33].

In the limiting case where Δt is equivalent to the sampling rate of the data, the KM coefficients take the form

$$\mathcal{M}^{[\ell,m]}(x_1, x_2) = \frac{1}{\Delta t} \langle \Delta y_1^\ell \Delta y_2^m |_{y_1(t)=x_1, y_2(t)=x_2} \rangle, \quad \Delta y_i = y_i(t + \Delta t) - y_i(t). \quad (3)$$

The algebraic relations between the KM coefficients and functions in Eq. (2) are given by [19,32]

$$\mathcal{M}^{[1,0]} = N_1, \quad (4)$$

$$\mathcal{M}^{[0,1]} = N_2,$$

$$\begin{aligned} \mathcal{M}^{[1,1]} &= g_{1,1}g_{2,1} + g_{1,2}g_{2,2}, \\ \mathcal{M}^{[2,0]} &= [g_{1,1}^2 + g_{1,2}^2], \\ \mathcal{M}^{[0,2]} &= [g_{2,1}^2 + g_{2,2}^2]. \end{aligned} \quad (5)$$

An explicit derivation can be found in Appendix A. Evidently, this underdetermined set of five equations is insufficient to uncover the six functions of a general stochastic diffusion process. One must bare this in mind, for the same issue will arise again when reconstructing jump-diffusion processes from data. Nonetheless, under certain assumptions it is possible to reduce the dimension of the problem and therefore obtain a system of equation which is not underdetermined. Two methods for these cases are presented in Ref. [19] and another criterion will be presented later.

In order to relate the results obtained from studying the KM coefficients against the theoretical functions, we propose a method to assess the difference between the values of the theoretically expected functions and the estimated values of the KM coefficients. Since for bivariate processes the KM coefficients are two-dimensional—as are the parameters of Eq. (1)—an adequate “distance” measure between the resulting two-dimensional surfaces is required.

Following Ref. [20], we propose a distance measure that allows for the variability of the density of data in some regions of the underlying space to be taken into consideration.

Let $f^{[\ell,m]}(y_1, y_2)$ denote the theoretical value for orders (ℓ, m) introduced in the model, i.e., a nonlinear combination of the various parameters of the system. The distance between each surface can be defined as

$$\int \int_U (\mathcal{M}^{[\ell,m]}(y_1, y_2) - f^{[\ell,m]}(y_1, y_2))^2 dy_1 dy_2 =: V^2, \quad (6)$$

where U denotes the domain of $\mathcal{M}^{[\ell,m]}(y_1, y_2)$. The least-squared distance volume V between the surfaces is zero if $\mathcal{M}^{[\ell,m]}(y_1, y_2) = f^{[\ell,m]}(y_1, y_2)$. It is this volume that one aims to minimize such that the reconstructed KM coefficients match the underlying theoretical functions in the model. Since $\mathcal{M}^{[\ell,m]}(y_1, y_2)$ is a real-valued function measured over a distribution space U , the density of data points is not uniform over U . This implies that a comparative measure on distances between $\mathcal{M}^{[\ell,m]}(y_1, y_2)$ and $f^{[\ell,m]}(y_1, y_2)$ would be non-normalized to the density of points of the space. We therefore introduce a normalization to Eq. (6) that ensures the less dense areas of U are normalized accordingly, thus

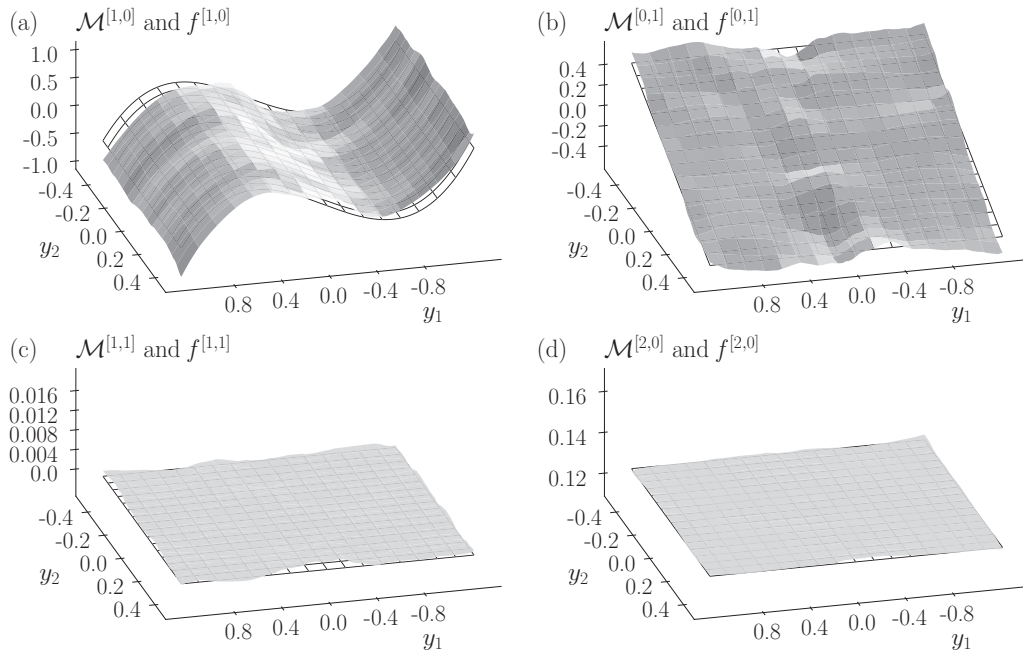


FIG. 1. Two-dimensional Kramers-Moyal coefficients $\mathcal{M}^{[\ell,m]}$ for two independent diffusion processes given by Eq. (11). The uncovered KM surfaces match the expectation. In panel (a) the cubic term in the drift term $N_1 = -x_1^3 + x_1$ along the first variable is visible in $\mathcal{M}^{[1,0]}$, and in panel (b) the negative-slope surface is visible in $\mathcal{M}^{[0,1]}$. In panel (d) the flat surfaces reproduce as well the expected form of the constant terms involved in the diffusion terms for $\mathcal{M}^{[2,0]}$. Moreover, in panel (c), $\mathcal{M}^{[1,1]}$, which accounts for the stochastic coupling terms of all diffusion terms, is also zero almost everywhere, as expected, given that $g_{1,2}$ and $g_{2,1}$ are zero. In each panel, the theoretical expected surface, given by Eqs. (4) and (6), is indicated by a grid, with $f^{[\ell,m]}$ denoting the respective theoretical values introduced in the model.

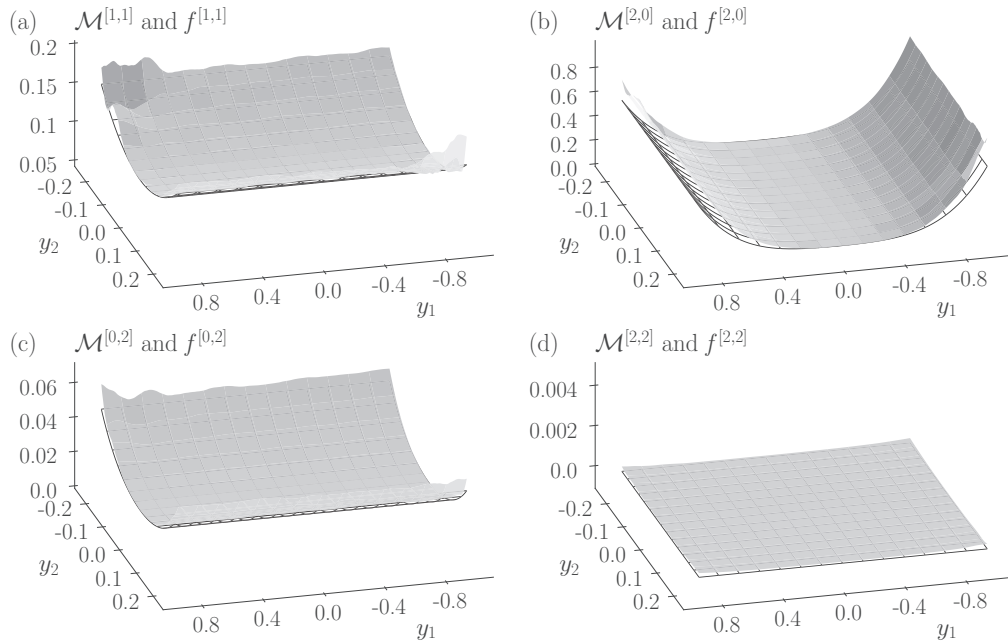


FIG. 2. Two-dimensional Kramers-Moyal coefficients $\mathcal{M}^{[l,m]}$ for two independent diffusion processes given by Eq. (12) and the theoretical expected functions $f^{[l,m]}$ associated with each coefficient, according to Eq. (6). KM coefficients $\mathcal{M}^{[1,1]}$, $\mathcal{M}^{[2,0]}$, $\mathcal{M}^{[0,2]}$, and $\mathcal{M}^{[2,2]}$ exhibit the quadratic multiplicative dependencies of the diffusion terms. In addition, $\mathcal{M}^{[1,1]}$ in panel (a) displays both an offset from zero as well as a quadratic shape, entailing the desired results emerging from Eq. (6), i.e., the noise-coupling term $g_{1,2}$ with $g_{2,2}$. In panel (b) $\mathcal{M}^{[2,0]}$ displays an offset and has a minimum close to $g_{1,1}^2/2 + g_{1,2}^2/2 = 0.13$. We also show the higher-order coefficient $\mathcal{M}^{[2,2]}$ in panel (d) and the corresponding theoretically expected value [given by Eq. (6)], both of which vanish. All obtained KM surfaces fit considerably well their theoretically expected ones ($V_{\text{err}}^{[1,1]} = 0.03$, $V_{\text{err}}^{[2,0]} = 0.94$, $V_{\text{err}}^{[0,2]} = 0.03$, $V_{\text{err}}^{[2,2]} < 0.01$; error volumes are estimated over the displayed domain).

mitigating the effect of scarcity of points at the borders of U and an overestimation of V due to outliers in the distribution. We derive such a normalization by considering the zeroth-order KM coefficient $\mathcal{M}^{[0,0]}(y_1, y_2)$ which captures exactly the density of points in U , although it is in itself not normalized as a distribution. The resulting normalized volume error measure V_{err} between surfaces takes the form (state dependencies not explicit)

$$\int \int_U (\mathcal{M}^{[l,m]} - f^{[l,m]})^2 p(y_1, y_2) dy_1 dy_2 = V_{\text{err}}^2, \quad (7)$$

where $p(\cdot)$ denotes the probability density. Coincidentally, the numerical evaluation implemented via either a histogram or a kernel-based estimator immediately yields this density, i.e., the zeroth power of the right-hand side of Eq. (3), before applying the estimation operator. This makes it easy to retrieve $p(y_1, y_2)$ as one numerically evaluates data.

With this at hand, it is now possible to relate theoretical and numerical results and to quantify the deviation of the obtained KM coefficients from the functions employed.

To showcase what two-dimensional KM coefficients are as well as how to identify drift and diffusion terms of bivariate diffusion processes, we present in the following two exemplary processes with *a priori* known coefficients. In this manner, by employing Eqs. (4) and (5), one can judge the outcome of the KM coefficient estimation procedure from discrete data in comparison with the expected theoretical functions.

We begin with two uncoupled processes, where one has constant diffusion and a quartic potential as the drift term:

$$\begin{aligned} \mathbf{N} &= \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} = \begin{pmatrix} -x_1^3 + x_1 \\ -x_2 \end{pmatrix}, \\ \mathbf{g} &= \begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix}. \end{aligned} \quad (8)$$

In Fig. 1, we show the corresponding KM coefficients $\mathcal{M}^{[1,0]}$, $\mathcal{M}^{[0,1]}$, $\mathcal{M}^{[1,1]}$, and $\mathcal{M}^{[2,0]}$ together with the theoretically expected functions. The per-design cubic-linear function ($N_1 = -x_1^3 + x_1$) acting as drift coefficient along the first dimension as well as the negatively sloped surface of $N_2 = -x_2$ are evident. Likewise, the constant diffusion term leads to a flat constant-valued $\mathcal{M}^{[2,0]}$, and the absence of any nondiagonal elements ($g_{1,2} = g_{2,1} = 0$) agrees with the zero-valued $\mathcal{M}^{[1,1]}$. Alongside the surfaces are plotted the theoretically expected values, which agree well with the data recovery.

We next extend Eq. (8) by adding multiplicative noise to the diffusion term and by including a noise-coupling term $g_{1,2} \neq 0$:

$$\begin{aligned} \mathbf{N} &= \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} = \begin{pmatrix} -x_1^3 + x_1 \\ -x_2 \end{pmatrix}, \\ \mathbf{g} &= \begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix} = \begin{pmatrix} 0.1 + x_1^2 & 0.5 \\ 0.0 & 0.2 + 2x_2^2 \end{pmatrix}. \end{aligned} \quad (9)$$

The recovered KM coefficients (see Fig. 2) of the drift terms remain unaltered, but as posited the second-order KM

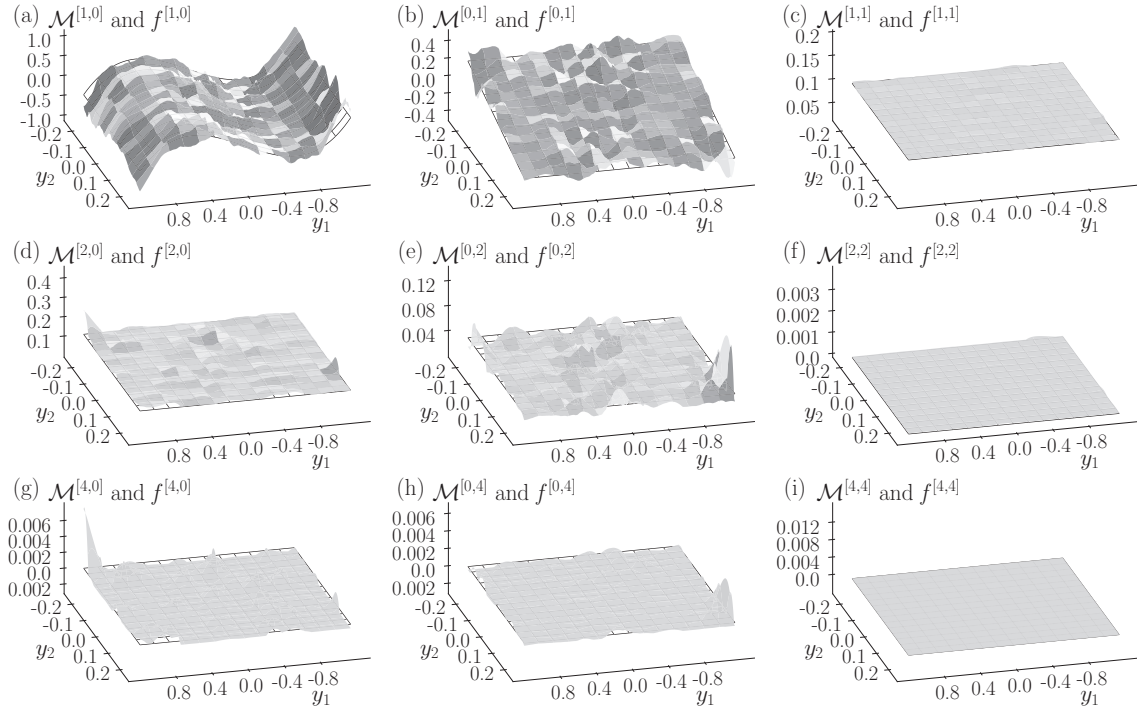


FIG. 3. Two-dimensional Kramers-Moyal coefficients $\mathcal{M}^{[\ell,m]}$ of bivariate diffusion processes given by Eq. (15) (with $\phi = 0.0$) together with the theoretically expected functions $f^{[\ell,m]}$ associated with each coefficient according to Eqs. (13), (15), and (18). KM coefficients $\mathcal{M}^{[1,0]}$, $\mathcal{M}^{[0,1]}$, $\mathcal{M}^{[1,1]}$, $\mathcal{M}^{[2,0]}$, $\mathcal{M}^{[0,2]}$, $\mathcal{M}^{[2,2]}$, $\mathcal{M}^{[4,0]}$, $\mathcal{M}^{[0,4]}$, and $\mathcal{M}^{[4,4]}$ are shown in panels (a)–(i), respectively. Although seemingly small, the higher-order moments [panels (f)–(i)] are all present and nonzero. We find $\mathcal{M}^{[4,0]} = 0.012$ in panel (g) and $\mathcal{M}^{[0,4]} = 0.009$ in panel (h), as expected from Eq. (18). All obtained KM surfaces fit considerably well their theoretically expected ones ($V_{\text{err}}^{[1,0]} = 0.68$, $V_{\text{err}}^{[0,1]} = 0.18$, $V_{\text{err}}^{[1,1]} < 0.01$, $V_{\text{err}}^{[2,0]} = 0.01$, $V_{\text{err}}^{[0,2]} = 0.01$, $V_{\text{err}}^{[2,2]} < 0.01$, $V_{\text{err}}^{[4,0]} = 0.01$, $V_{\text{err}}^{[0,4]} = 0.01$, $V_{\text{err}}^{[4,4]} < 0.01$; error volumes are estimated over the displayed domain).

coefficients, i.e., $\mathcal{M}^{[1,1]}$, $\mathcal{M}^{[2,0]}$, $\mathcal{M}^{[0,2]}$, and $\mathcal{M}^{[2,2]}$, clearly exhibit the influence of the multiplicative noise. The quadratic multiplicative dependencies of $\mathcal{M}^{[2,0]}$ and $\mathcal{M}^{[0,2]}$ and their offsets from zero are evident. More pertinently, one can notice $\mathcal{M}^{[1,1]}$ to display the expected shape arising from Eq. (5), i.e., this value is nonzero and exhibits the parabolic shape of $g_{1,2}g_{2,2} = 0.5(0.2 + 2x_2^2)$. For $x_2 = 0$, the minimum of $\mathcal{M}^{[1,1]}$ coincides with 0.1, as expected. This indicates that the presence of the multiplicative noise does not hinder the assertion of the KM coefficients. Again, the recovered KM coefficients match the theoretical ones.

IV. BIVARIATE JUMP-DIFFUSION PROCESSES

The KM coefficients of bivariate jump-diffusion processes take the following form [under the parameter prescription used in the jump-diffusion model in Eq. (1)] [19,32,33]:

$$\begin{aligned}\mathcal{M}^{[1,0]} &= N_1, \\ \mathcal{M}^{[0,1]} &= N_2,\end{aligned}\tag{10}$$

$$\begin{aligned}\mathcal{M}^{[1,1]} &= g_{1,1}g_{2,1} + g_{1,2}g_{2,2}, \\ \mathcal{M}^{[2,0]} &= [g_{1,1}^2 + s_{1,1}\lambda_1 + g_{1,2}^2 + s_{1,2}\lambda_2], \\ \mathcal{M}^{[0,2]} &= [g_{2,1}^2 + s_{2,1}\lambda_1 + g_{2,2}^2 + s_{2,2}\lambda_2],\end{aligned}\tag{11}$$

$$\begin{aligned}\mathcal{M}^{[2,2]} &= [s_{1,1}s_{2,1}\lambda_1 + s_{1,2}s_{2,2}\lambda_2], \\ \mathcal{M}^{[4,0]} &= 3[s_{1,1}^2\lambda_1 + s_{1,2}^2\lambda_2],\end{aligned}\tag{12}$$

$$\begin{aligned}\mathcal{M}^{[0,4]} &= 3[s_{2,1}^2\lambda_1 + s_{2,2}^2\lambda_2], \\ \mathcal{M}^{[4,4]} &= 9[s_{1,1}^2s_{2,1}^2\lambda_1 + s_{1,2}^2s_{2,2}^2\lambda_2], \\ \mathcal{M}^{[6,0]} &= 15[s_{1,1}^3\lambda_1 + s_{1,2}^3\lambda_2], \\ \mathcal{M}^{[0,6]} &= 15[s_{2,1}^3\lambda_1 + s_{2,2}^3\lambda_2],\end{aligned}\tag{13}$$

$$\begin{aligned}\mathcal{M}^{[6,6]} &= 225[s_{1,1}^3s_{2,1}^3\lambda_1 + s_{1,2}^3s_{2,2}^3\lambda_2], \\ \mathcal{M}^{[8,0]} &= 105[s_{1,1}^4\lambda_1 + s_{1,2}^4\lambda_2], \\ \mathcal{M}^{[0,8]} &= 105[s_{2,1}^4\lambda_1 + s_{2,2}^4\lambda_2], \\ \mathcal{M}^{[8,8]} &= 11\,025[s_{2,1}^4\lambda_1 + s_{2,2}^4\lambda_2],\end{aligned}\tag{14}$$

where $\langle \xi_{ij}^{2\ell} \rangle = s_{ij}^\ell$ are the variances of the Gaussian-distributed jump amplitudes. An extended derivation can be found in Appendix A. The last equations here are taken from the general form

$$M^{[2\ell, 2m]} = \frac{(2\ell)! (2m)!}{2^\ell \ell! 2^m m!} [s_{1,1}^\ell s_{2,1}^m \lambda_1 + s_{1,2}^\ell s_{2,2}^m \lambda_2].$$

A. Understanding the impact of jumps

As an illustrative case study, we investigate a general jump-diffusion process that is based on Eq. (9) but excludes the multiplicative diffusion terms. Taking into account the effect of the jump terms, but maintaining the system independent in at least one of the dimensions, we extend Eq. (9) to include jumps only in the diagonal terms of ξ :

$$\begin{aligned} N &= \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} = \begin{pmatrix} -x_1^3 + x_1 \\ -x_2 \end{pmatrix}, \\ g &= \begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix} = \begin{pmatrix} 0.1 & 0.5 \\ 0.0 & 0.2 \end{pmatrix}, \\ \xi &= \begin{pmatrix} \xi_{1,1} & \xi_{1,2} \\ \xi_{2,1} & \xi_{2,2} \end{pmatrix} = \begin{pmatrix} 0.2 & 0.0 \\ \phi & 0.1 \end{pmatrix}, \\ \lambda &= \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.3 \end{pmatrix}, \end{aligned} \quad (15)$$

where for the present case $\phi = 0.0$. In this manner, jumps are added to the first dimension of the process, having an amplitude of $\xi_{1,1} = 0.2$ and occurring every $0.1t$, given $\lambda_1 = 0.1$. Similarly, jumps are added to the second dimension, $\xi_{2,2} = 0.1$, but the jumps occur three times more often than the aforementioned, given $\lambda_2 = 0.3$. The influence of jumps can be observed across all KM coefficients (see Fig. 3). The previously smooth KM surfaces become rugged from the fast variations emerging due to the jumps, and the higher-order KM coefficients—although small compared to the lower-order ones—clearly do not vanish. This indicates that the continuous stochastic modeling of time series of complex systems (the white-noise-driven Langevin equation) is invalid for jump-diffusion processes. Modeling these processes with only the first two orders of the KM expansion of the master equation is therefore insufficient.

In order to understand further if it is possible to uncover the jump amplitude terms of coupled processes, we use the previous model Eq. (15) with $\phi = 0.3$, thereby effectively introducing a stochastic coupling via the off-diagonal elements of the jump matrix ξ . We show, in Fig. 4, the corresponding fourth-order KM coefficients. The impact of the stochastic coupling is visible, although small, in $\mathcal{M}^{[4,4]}$, which is no longer zero. Likewise, $\mathcal{M}^{[4,0]}$ and $\mathcal{M}^{[0,4]}$ also do not vanish. In Appendix B, we present the corresponding KM coefficients up to order 8.

B. Criteria for recovering coefficients in diffusion and jump-diffusion models

For the case of vanishing off-diagonal elements $g_{2,1}$ and $\xi_{1,2}$, we can identify ways to recover the remaining coefficients of these processes.

First, given that the noise $d\omega$ is Gaussian distributed, g is sign-reversal symmetric and one can thus assume that it takes only positive values. One obtains that if $\mathcal{M}^{[1,1]} = 0$ then at least two elements of g must be zero, and if $\mathcal{M}^{[2,2]} = 0$ then at least two elements of ξ must be zero (by assuming that λ_1 and λ_2 are nonvanishing rates). These findings reduce the dimensionality of the estimation procedure and ensure that the underlying processes are less complex than the full-fledged description of Eq. (1), although they do not grant which coefficients are zero valued.

Second, if one either employs a heuristic argument of independence of the jump processes or neglects the off-diagonal jump amplitudes $\xi_{1,2}$ and $\xi_{2,1}$ (e.g., by assuming they are small compared to the diagonal terms of ξ), one finds the following approximations:

$$\begin{aligned} \frac{1}{5} \frac{\mathcal{M}^{[6,0]}}{\mathcal{M}^{[4,0]}} &= \frac{1}{5} \frac{15}{3} \frac{s_{1,1}^3 \lambda_1}{s_{1,1}^2 \lambda_1} = s_{1,1}, \\ \frac{1}{5} \frac{\mathcal{M}^{[0,6]}}{\mathcal{M}^{[0,4]}} &= \frac{1}{5} \frac{15}{3} \frac{s_{2,2}^3 \lambda_2}{s_{2,2}^2 \lambda_2} = s_{2,2}. \end{aligned} \quad (16)$$

Likewise, the jump rates λ_1 and λ_2 can be obtained equivalently as

$$\begin{aligned} \frac{105}{9} \frac{\mathcal{M}^{[4,0]^2}}{\mathcal{M}^{[8,0]}} &= \frac{105}{9} \frac{3^2}{105} \frac{(s_{1,1}^2 \lambda_1)^2}{s_{1,1}^4 \lambda_1} = \lambda_1, \\ \frac{105}{9} \frac{\mathcal{M}^{[0,4]^2}}{\mathcal{M}^{[0,8]}} &= \frac{105}{9} \frac{3^2}{105} \frac{(s_{2,2}^2 \lambda_2)^2}{s_{2,2}^4 \lambda_2} = \lambda_2. \end{aligned} \quad (17)$$

Taking again model Eq. (15) with $\phi = 0.0$ and following Eq. (16), we obtain

$$\begin{aligned} s_{1,1}^{\text{est}} &= 0.16 \approx 0.2 = s_{1,1}, \\ s_{2,2}^{\text{est}} &= 0.09 \approx 0.1 = s_{2,2}. \end{aligned}$$

These estimated values (indicated by the superscript “est”) are close to the actual ones. The criteria and approximations are especially relevant when constructing or analyzing systems which are known to have a specific unidirectional stochastic coupling form, e.g., a master-slave system, where, for example, the noise or the slave system is dictated by the driving master system.

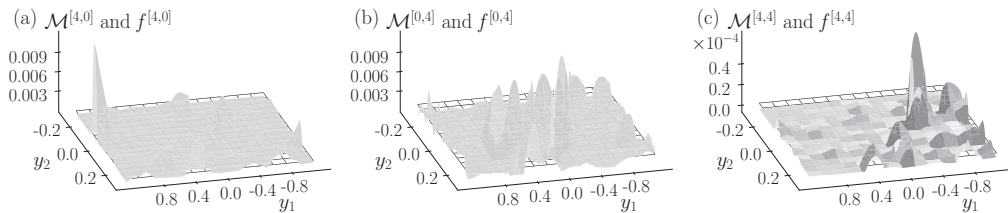


FIG. 4. Two-dimensional Kramers-Moyal coefficients $\mathcal{M}^{[4,0]}$, $\mathcal{M}^{[0,4]}$, and $\mathcal{M}^{[4,4]}$ of bivariate jump-diffusion processes given by Eq. (15) (with $\phi = 0.3$) together with the respective theoretically expected functions f , associated with each coefficient according to Eqs. (13), (15), and (18). Notice that the estimated KM coefficients agree well with the theoretical expected functions in all orders. For further details, see Appendix B.

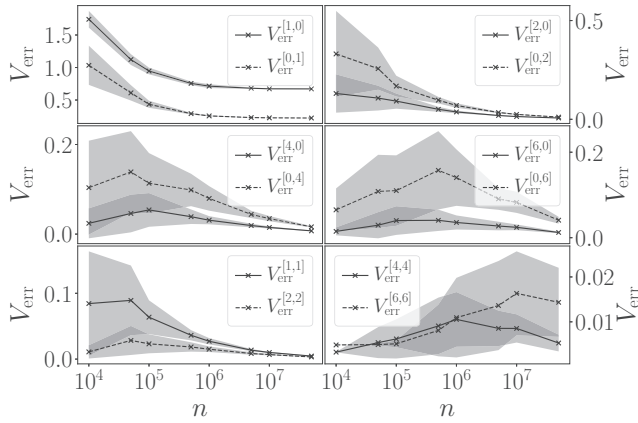


FIG. 5. Error volume V_{err} for a bivariate jump-diffusion process [Eq. (15)] depending on the number of data points n in the time series, with the abscissa given in logarithmic scale. Each process is numerically integrated with random initial conditions, for varying number of data points $n \in [10^4, 5 \times 10^7]$ and over 50 times. The time sampling used was of 10^{-3} . The average value of V_{err} and one standard deviation (shaded area) are displayed. Notice the clear decrease on all KM coefficients with either $\ell = 0$ or $m = 0$, e.g., $\mathcal{M}^{[2,0]}$ or $\mathcal{M}^{[0,2]}$, as the number of data points n increases. This can be seen since the volume between the theoretically expected values and the KM coefficients decreases consistently, i.e., $V_{\text{err}}^{[\ell,m]}$ decreases for an increasing number of data points. The KM coefficients with $\ell \neq 0$ and $m \neq 0$, such as $\mathcal{M}^{[4,4]}$ or $\mathcal{M}^{[6,6]}$, present themselves as nondecreasing, but the error volume is overall considerably small in value (cf. Fig. 6). It is important to notice that $V_{\text{err}}^{[1,0]}$ does not converge to zero since the KM coefficient is associated with the quartic potential (i.e., the term $N_1 = -x^3 + x$). Due to its shape, the process has two preferred states, at either $x = -1$ or 1 , and thus spends little time at any intermediary point, like $x = 0$, damaging the statistics of the recovery.

C. Factors influencing the quality of recovery of coefficients

In order to validate the quality of the nonparametric recovery of the KM coefficients, we now turn to two critical aspects: first, bivariate processes may require a high number of data points in a time series for the estimation to be reliable; second, the interplay between the drift, diffusion, and jump parts of a stochastic processes may render the estimation incorrect.

Addressing these aspects, we include a more contrived model involving stochastic couplings and interactions in both the diffusion and jump terms, thus theoretically resulting in having all higher-order KM coefficients nonzero, and especially the KM coefficients with $\ell \neq 0$ and $m \neq 0$. The parameters for the model read

$$\begin{aligned} N &= \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} = \begin{pmatrix} -x_1^3 + x_1 \\ -x_2 \end{pmatrix}, \\ g &= \begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix} = \begin{pmatrix} 0.1 & 0.5 \\ \alpha & 0.2 \end{pmatrix}, \\ \xi &= \begin{pmatrix} \xi_{1,1} & \xi_{1,2} \\ \xi_{2,1} & \xi_{2,2} \end{pmatrix} = \begin{pmatrix} 0.2 & 0.5 \\ \beta & 0.1 \end{pmatrix}, \\ \lambda &= \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.3 \end{pmatrix}. \end{aligned} \quad (18)$$

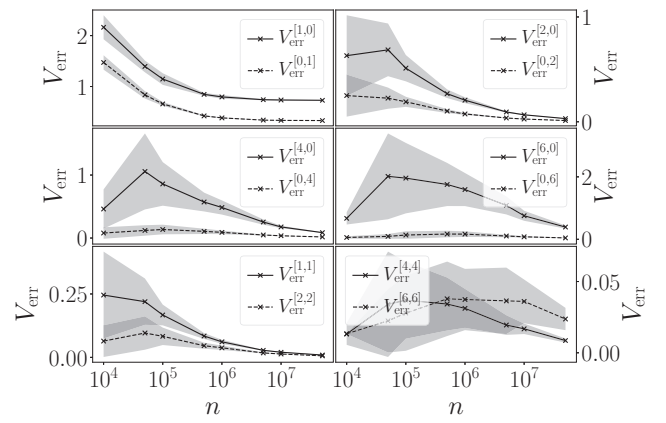


FIG. 6. Same as Fig. 5 but for the bivariate jump-diffusion process [Eq. (18)] with $\alpha = \beta = 0.3$, with the abscissa given in logarithmic scale. Integration parameters are as in Fig. 5.

The diffusion-scaling parameter α and the jump-scaling parameter β (jump term) can be freely varied.

Let us focus first on the number of data points in a time series. We utilize the models Eq. (15) with $\phi = 0.3$ and Eq. (18) with $\alpha = \beta = 0.3$, and show in Figs. 5 and 6, respectively, the error volumes V_{err} for the KM coefficients for an increasing number of data points. The reliability of the recovery of the KM coefficient is valid for a higher amount of data ($n \geq 10^5$), as expected, although the scarcity of data poses no extensive problem for the calculation. It is especially important to notice that a time series with a lower amount of data entails naturally fewer jumps in the process, hindering the possibility of accurately recovering the jump terms from such short time series. For $n \geq 10^6$, the estimation seems reliable, the standard deviations become minute, and most error values approach zero, i.e., the theoretical and estimated KM surfaces are close. Such a large number of data points might not be available when investigating time-varying dynamical (e.g., biological) systems. Nevertheless, the amount of data needed to reliably estimate KM coefficients can be considerably reduced with kernel-based estimators [40].

One remark is necessary on the recovery of the drift terms. The presence of noise and jumps in the process takes its toll on the recovery of the exact form of the KM coefficients as well as the explicit dependence of the state variables, i.e., the quartic potential in both Eqs. (15) and (18). A finer time sampling can help to improve the results.

To further test the limitation of retrieving the KM coefficients from data, we utilize model Eq. (18) once more and investigate the influence of the diffusion-scaling parameter α and the jump-scaling parameter β . For increasing diffusion-scaling parameter α ($\alpha \in [10^{-2}, 10^2]$) and jump-scaling parameter $\beta = 0.3$, we observe a considerable impact on the error volume V_{err} after the order of magnitude on the diffusion parameter α is tenfold bigger in comparison to the diffusion parameter $g_{1,2}$ (Fig. 7). Similarly, for increasing jump-scaling parameter β ($\beta \in [10^{-2}, 10^2]$) and diffusion-scaling parameter $\alpha = 0.3$, the error volume V_{err} is considerably impacted already when β is of similar size as the other parameters, namely, $\xi_{1,2} = 0.5$ (Fig. 8).

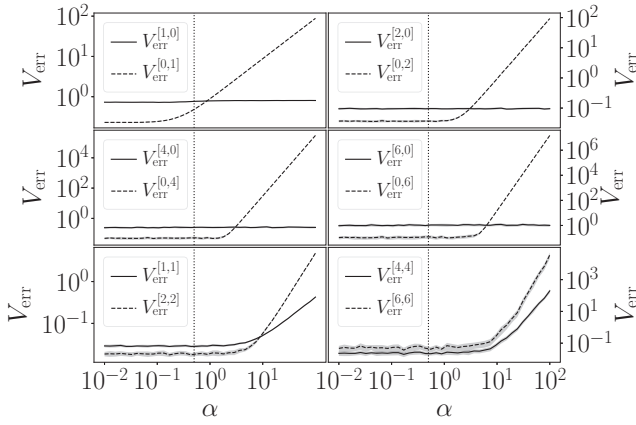


FIG. 7. Error volume V_{err} for the bivariate jump-diffusion process Eq. (18) for a varying diffusion-scaling parameter $\alpha \in [10^{-2}, 10^2]$, given in double logarithmic scale. The vertical dotted line at $\alpha = 0.5$ indicates the point where the diffusion-scaling parameter $\alpha = g_{2,1} = 0.5$ is equal to $g_{1,2} = 0.5$. A small value of the diffusion-scaling parameter α , in comparison to the diffusion parameters $g_{2,1}$ and $g_{1,2}$, ensures a good reconstruction, i.e., a small error volume V_{err} . The average and one-standard deviations (shaded area) are displayed. For each point 50 iterations are taken, each with a total number of data points of 5×10^6 and a time sampling of 10^{-3} .

These findings point to the difficulty of recovering the KM coefficients in the presence of jumps. Nonetheless, our findings indicate that the current understanding, modeling, and numerical recovery of KM surfaces, for the case of jumps of comparable size to the diffusion terms, is possible and reliable [41]. This can be performed in minimal times on a regular computer [42].

V. CONCLUSION

We introduced the bivariate jump-diffusion process, which consists of two-dimensional diffusion and two-dimensional jumps that can be coupled to one another.

For such a process we presented a data-driven, nonparametric estimation procedure of higher-order Kramers-Moyal coefficients and investigated its pros and cons using synthetic bivariate time series from continuous and discontinuous processes. The procedure allows one to reconstruct relevant aspects of the underlying jump-diffusion processes and to recover the underlying parameters.

Having now a traceable mathematical framework, the model can be extended to embody other noise and jump properties. An extension from the underlying Wiener process to include, e.g., fractional Brownian motion is straightforward

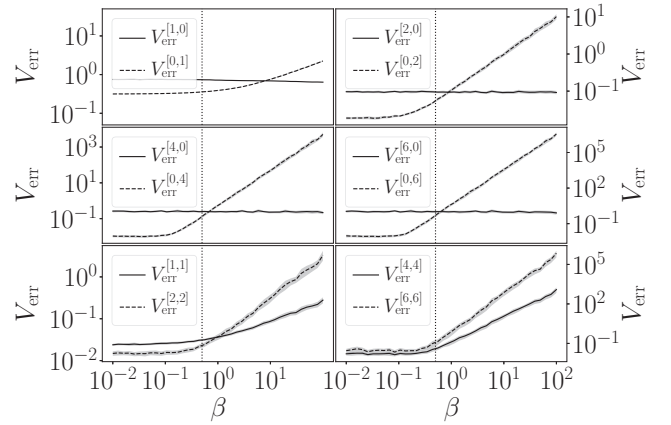


FIG. 8. Error volume V_{err} for the bivariate jump-diffusion process Eq. (18) for a varying jump-scaling parameter $\beta \in [10^{-2}, 10^2]$, given in double logarithmic scale. The vertical dotted line indicates the biggest jump term value $\xi_{1,2} = 0.5$ to compare with β . In direct analogy to Fig. 7, a small jump-scaling parameter β ensures a good reconstruction, i.e., a small error volume V_{err} . Increasing values of the jump-scaling parameter β in comparison to the other parameters in the system make the reconstruction unreliable. The iteration scheme is identical to the one in Fig. 7.

ward but nevertheless requires further investigations to derive an explicit forward Kolmogorov equation [19]. Also, a generalization to continuous jump processes—originating from alpha-stable or other heavy-tailed distributions (the Lévy noise-driven Langevin equation)—is possible, however, with the drawback that calculating the conditional moments may not always be mathematically possible [19]. On the other hand, a numerical estimation of generalized moments should be possible but these still require a physical interpretation.

We are confident that our approach provides a general avenue to further understanding of interacting complex systems (e.g., brain or power grids [33,43–45]) the dynamics of which exhibit nontrivial noise contributions.

ACKNOWLEDGMENTS

The authors would like to thank Thorsten Rings and Mehrnaz Anvari for interesting discussions, and Francisco Meirinhos for the help in devising the methodology behind the results. L.R.G. thanks Giulia di Nunno and Dirk Witthaut for the support, and gratefully acknowledges support by Grant No. VH-NG-1025 and a scholarship from the E.ON Stipendiefonds.

APPENDIX A: EXTENDED DERIVATION OF THE TWO-DIMENSIONAL KRAMERS-MOYAL COEFFICIENTS FOR A JUMP-DIFFUSION PROCESS

The following derivations stem from Eq. (3) and apply to the two-dimensional jump-diffusion process (y_1, y_2) , as in Eq. (1). All orders of the Kramers-Moyal coefficients $\mathcal{M}^{[\ell, m]}$ are $(\ell, m) \in \mathbb{N}_+$.

1. Kramers-Moyal coefficients $\mathcal{M}^{[1,0]}$ and $\mathcal{M}^{[0,1]}$

$$\begin{aligned}
\mathcal{M}^{[1,0]}(x_1, x_2) &= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (dy_1)^1 (dx_2)^0 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle dy_1 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle N_1 dt + g_{1,1} dw_1 + g_{1,2} dw_2 + \xi_{1,1} dJ_1 + \xi_{1,2} dJ_2 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} [N_1 dt + g_{1,1} \langle dw_1 \rangle + g_{1,2} \langle dw_2 \rangle + \langle \xi_{1,1} \rangle \langle dJ_1 \rangle + \langle \xi_{1,2} \rangle \langle dJ_2 \rangle] \\
&= N_1,
\end{aligned}$$

where $\langle g_{i,j} dW_j \rangle = \langle g_{i,j} \rangle \langle dW_j \rangle = 0$, because a Wiener process has the property $\langle dW_j \rangle = 0$. Further, $\langle \xi_{i,j} dJ_j \rangle = \langle \xi_{i,j} \rangle \langle dJ_j \rangle = 0$, since $\xi_{i,j}$ is a Gaussian with zero mean, i.e., $\langle \xi_{i,j} \rangle = 0$.

The same is true, mutatis mutandis-, for $\mathcal{M}^{[0,1]}$.

2. Kramers-Moyal coefficient $\mathcal{M}^{[1,1]}$

$$\begin{aligned}
\mathcal{M}^{[1,1]} &= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (dy_1)^1 (dy_2)^1 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (N_1 dt + g_{1,1} dw_1 + g_{1,2} dw_2 + \xi_{1,1} dJ_1 + \xi_{1,2} dJ_2) \\
&\quad \times (N_2 dt + g_{2,1} dw_1 + g_{2,2} dw_2 + \xi_{2,1} dJ_1 + \xi_{2,2} dJ_2) \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \left[N_1 N_2 dt + g_{1,1} g_{2,1} \langle (dw_1)^2 \rangle \frac{1}{dt} + g_{1,2} g_{2,2} \langle (dw_2)^2 \rangle \frac{1}{dt} + O(dt) \right] \\
&= g_{1,1} g_{2,1} + g_{1,2} g_{2,2},
\end{aligned}$$

where higher-order terms $O(dt)^\epsilon$, with $\epsilon > 0$, vanish in the limit $dt \rightarrow 0$. Recall as well $\langle (dw_i)^2 \rangle = dt$.

3. Kramers-Moyal coefficients $\mathcal{M}^{[2,0]}$ and $\mathcal{M}^{[0,2]}$

$$\begin{aligned}
\mathcal{M}^{[2,0]} &= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (dy_1)^2 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (N_1 dt + g_{1,1} dw_1 + g_{1,2} dw_2 + \xi_{1,1} dJ_1 + \xi_{1,2} dJ_2)^2 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \left[N_1^2 dt + g_{1,1}^2 \langle (dw_1)^2 \rangle \frac{1}{dt} + g_{1,2}^2 \langle (dw_2)^2 \rangle \frac{1}{dt} + \langle \xi_{1,1}^2 \rangle \langle (dJ_1)^2 \rangle \frac{1}{dt} + \langle \xi_{1,2}^2 \rangle \langle (dJ_2)^2 \rangle \frac{1}{dt} + O(dt) \right] \\
&= [g_{1,1}^2 + s_{1,1} \lambda_1 + g_{1,2}^2 + s_{1,2} \lambda_2],
\end{aligned}$$

using the previously employed nomenclature $\langle \xi_{ij}^2 \rangle = \sigma_{\xi_{ij}}^2 = s_{ij}$ as well as $\langle (dJ_i)^2 \rangle = \lambda_i dt$.

Mutatis mutandis, the case for $\mathcal{M}^{[0,2]}$ reads as

$$\mathcal{M}^{[0,2]} = [g_{2,1}^2 + s_{2,1} \lambda_1 + g_{2,2}^2 + s_{2,2} \lambda_2].$$

4. Kramers-Moyal coefficient $\mathcal{M}^{[2,2]}$

$$\begin{aligned}
\mathcal{M}^{[2,2]} &= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (dy_1)^2 (dy_2)^2 \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (N_1 dt + g_{1,1} dw_1 + g_{1,2} dw_2 + \xi_{1,1} dJ_1 + \xi_{1,2} dJ_2)^2 \\
&\quad \times (N_2 dt + g_{2,1} dw_1 + g_{2,2} dw_2 + \xi_{2,1} dJ_1 + \xi_{2,2} dJ_2)^2 \rangle|_{y_1(t)=x_1, y_2(t)=x_2}
\end{aligned}$$

$$\begin{aligned}
&= \lim_{dt \rightarrow 0} \frac{1}{dt} [\text{terms}(N_1, N_2, O(dt^4)) + \text{terms}(g_{ij}, O(dt^2)) + \text{terms}(\text{mixing } \xi_{ij}) \\
&\quad + \langle \xi_{1,1}^2 \rangle \langle \xi_{2,1}^2 \rangle \langle (dJ_1)^4 \rangle + \langle \xi_{1,2}^2 \rangle \langle \xi_{2,2}^2 \rangle \langle (dJ_2)^4 \rangle + \langle \xi_{1,1}^2 \rangle \langle \xi_{2,2}^2 \rangle \langle (dJ_1)^2 \rangle \langle (dJ_2)^2 \rangle + \langle \xi_{1,2}^2 \rangle \langle \xi_{2,1}^2 \rangle \langle (dJ_1)^2 \rangle \langle (dJ_2)^2 \rangle] \\
&= [s_{1,1}s_{2,1}\lambda_1 + s_{1,2}s_{2,2}\lambda_2].
\end{aligned}$$

Terms including dt on the right-hand side of the above equation vanish for $dt \rightarrow 0$, where as well $\langle \xi_{1,1}\xi_{1,2} \rangle = \langle \xi_{1,1} \rangle \langle \xi_{1,2} \rangle = 0$, and $\frac{1}{dt} [\langle (dJ_1)^2 \rangle \langle (dJ_2)^2 \rangle] = \frac{1}{dt} [\lambda_1 dt \lambda_2 dt] \propto dt$ vanishes in the limit $dt \rightarrow 0$.

5. Kramers-Moyal coefficients $\mathcal{M}^{[\ell,m]}$, for $2 \times (\ell, m) \geq 2$

For $(2\ell, 2m)$, with $(\ell, m) \geq 4$, the Kramers-Moyal coefficients $\mathcal{M}^{[2\ell, 2m]}$ are as follows:

$$\begin{aligned}
\mathcal{M}^{[2\ell, 2m]} &= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (dy_1)^{2\ell} (dy_2)^{2m} \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} \langle (N_1 dt + g_{1,1} dw_1 + g_{1,2} dw_2 + \xi_{1,1} dJ_1 + \xi_{1,2} dJ_2)^{2\ell} \\
&\quad \times (N_2 dt + g_{2,1} dw_1 + g_{2,2} dw_2 + \xi_{2,1} dJ_1 + \xi_{2,2} dJ_2)^{2m} \rangle|_{y_1(t)=x_1, y_2(t)=x_2} \\
&= \lim_{dt \rightarrow 0} \frac{1}{dt} [\langle \xi_{1,1}^{2\ell} \rangle \langle \xi_{2,1}^{2m} \rangle \langle (dJ_1)^{2(\ell+m)} \rangle + \langle \xi_{1,2}^{2\ell} \rangle \langle \xi_{2,2}^{2m} \rangle \langle (dJ_2)^{2(\ell+m)} \rangle] \\
&= [\langle \xi_{1,1}^{2\ell} \rangle \langle \xi_{2,1}^{2m} \rangle \lambda_1 + \langle \xi_{1,2}^{2\ell} \rangle \langle \xi_{2,2}^{2m} \rangle \lambda_2] \\
&= \frac{(2\ell)! (2m)!}{2^\ell \ell! 2^m m!} [s_{1,1}^\ell s_{2,1}^m \lambda_1 + s_{1,2}^\ell s_{2,2}^m \lambda_2].
\end{aligned}$$

In the last step, take the fact that the jump amplitudes $\xi_{i,j}$ are Gaussian distributed; thus, $\langle \xi_{i,j}^{2\ell} \rangle \propto \sigma_{\xi_{i,j}}^{2\ell} = s_{i,j}^\ell$. In this manner, all Kramers-Moyal coefficients $\mathcal{M}^{[2\ell, 2m]}$ with $(\ell, m) \geq 1$ are obtained.

APPENDIX B: EXTENDED RESULTS FOR MODELED DATA BY EQ. (15)

Figure 9 extends Fig. 4 and includes the Kramers-Moyal coefficients $\mathcal{M}^{[1,0]}$, $\mathcal{M}^{[0,1]}$, $\mathcal{M}^{[1,1]}$, $\mathcal{M}^{[2,0]}$, $\mathcal{M}^{[0,2]}$, $\mathcal{M}^{[2,2]}$, $\mathcal{M}^{[4,0]}$, $\mathcal{M}^{[0,4]}$, $\mathcal{M}^{[4,4]}$, $\mathcal{M}^{[6,0]}$, $\mathcal{M}^{[0,6]}$, $\mathcal{M}^{[6,6]}$, $\mathcal{M}^{[8,0]}$, and $\mathcal{M}^{[0,8]}$.

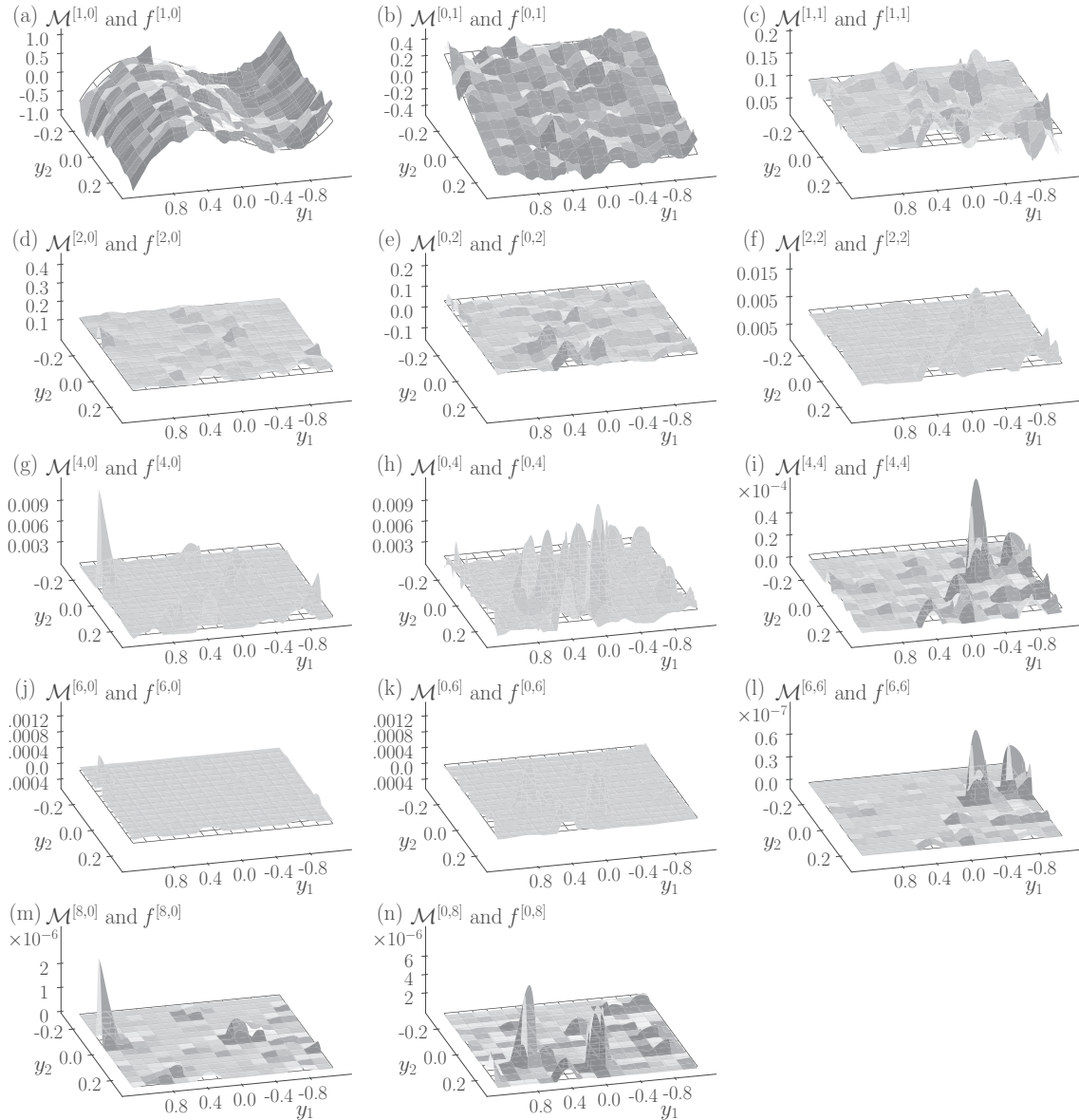


FIG. 9. Two-dimensional Kramers-Moyal coefficients $\mathcal{M}^{[l,m]}$ of bivariate jump-diffusion processes given by Eq. (15) and all theoretical expected functions $f^{[l,m]}$ associated with each KM coefficient according to Eqs. (13), (15), and (18). Shown are the KM coefficients $\mathcal{M}^{[1,0]}$, $\mathcal{M}^{[0,1]}$, $\mathcal{M}^{[1,1]}$, $\mathcal{M}^{[2,0]}$, $\mathcal{M}^{[0,2]}$, $\mathcal{M}^{[2,2]}$, $\mathcal{M}^{[4,0]}$, $\mathcal{M}^{[0,4]}$, $\mathcal{M}^{[4,4]}$, $\mathcal{M}^{[6,0]}$, $\mathcal{M}^{[0,6]}$, $\mathcal{M}^{[6,6]}$, $\mathcal{M}^{[8,0]}$, and $\mathcal{M}^{[0,8]}$. The respective error volumes read $V_{\text{err}}^{[1,0]} = 0.6836$, $V_{\text{err}}^{[0,1]} = 0.23$, $V_{\text{err}}^{[1,1]} = 0.01$, $V_{\text{err}}^{[2,0]} = 0.01$, $V_{\text{err}}^{[0,2]} = 0.03$, $V_{\text{err}}^{[2,2]} = 0.01$, $V_{\text{err}}^{[4,0]} = 0.02$, $V_{\text{err}}^{[0,4]} = 0.03$, $V_{\text{err}}^{[4,4]} < 0.01$, $V_{\text{err}}^{[6,0]} = 0.02$, $V_{\text{err}}^{[0,6]} = 0.01$, $V_{\text{err}}^{[6,6]} = 0.01$, $V_{\text{err}}^{[8,0]} = 0.04$, and $V_{\text{err}}^{[0,8]} = 0.26$.

-
- [1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Complex networks: Structure and dynamics, *Phys. Rep.* **424**, 175 (2006).
- [2] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, Synchronization in complex networks, *Phys. Rep.* **469**, 93 (2008).
- [3] E. T. Bullmore and D. S. Bassett, Brain graphs: Graphical models of the human brain connectome, *Annu. Rev. Clin. Psychol.* **7**, 113 (2011).
- [4] M. Barthélemy, Spatial networks, *Phys. Rep.* **499**, 1 (2011).
- [5] M. E. J. Newman, Communities, modules and large-scale structure in networks, *Nat. Phys.* **8**, 25 (2012).
- [6] P. Holme and J. Saramäki, Temporal networks, *Phys. Rep.* **519**, 97 (2012).
- [7] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, Multilayer networks, *J. Complex Netw.* **2**, 203 (2014).
- [8] A. S. Pikovsky, M. G. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences* (Cambridge University, Cambridge, England, 2001).

- [9] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. (Cambridge University, Cambridge, England, 2003).
- [10] E. Pereda, R. Quiñan Quiroga, and J. Bhattacharya, Nonlinear multivariate analysis of neurophysiological signals, *Prog. Neurobiol.* **77**, 1 (2005).
- [11] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis, *Phys. Rep.* **441**, 1 (2007).
- [12] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, Recurrence plots for the analysis of complex systems, *Phys. Rep.* **438**, 237 (2007).
- [13] K. Lehnertz, S. Bialonski, M.-T. Horstmann, D. Krug, A. Rothkegel, M. Staniek, and T. Wagner, Synchronization phenomena in human epileptic brain networks, *J. Neurosci. Methods* **183**, 42 (2009).
- [14] K. Lehnertz, Assessing directed interactions from neurophysiological signals: An overview, *Physiol. Meas.* **32**, 1715 (2011).
- [15] T. Stankovski, T. Pereira, P. V. E. McClintock, and A. Stefanovska, Coupling functions: Universal insights into dynamical interaction mechanisms, *Rev. Mod. Phys.* **89**, 045001 (2017).
- [16] H. Risken, *The Fokker-Planck Equation*, 2nd ed., Springer Series in Synergetics (Springer-Verlag, Berlin, 1996).
- [17] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
- [18] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, Approaching complexity by stochastic methods: From biological systems to turbulence, *Phys. Rep.* **506**, 87 (2011).
- [19] M. R. R. Tabar, *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems: Using the Methods of Stochastic Processes* (Springer, New York, 2019).
- [20] J. Prussek and K. Lehnertz, Measuring interdependences in dissipative dynamical systems with estimated Fokker-Planck coefficients, *Phys. Rev. E* **77**, 041914 (2008).
- [21] B. Lehle, Stochastic time series with strong, correlated measurement noise: Markov analysis in n dimensions, *J. Stat. Phys.* **152**, 1145 (2013).
- [22] T. Scholz, F. Rauschel, V. V. Lopes, B. Lehle, M. Wächter, J. Peinke, and P. G. Lind, Parameter-free resolution of the superposition of stochastic signals, *Phys. Lett. A* **381**, 194 (2017).
- [23] B. Wahl, U. Feudel, J. Hlinka, M. Wächter, J. Peinke, and J. A. Freund, Granger-causality maps of diffusion processes, *Phys. Rev. E* **93**, 022213 (2016).
- [24] P. G. Lind, M. Haase, F. Böttcher, J. Peinke, D. Kleinhans, and R. Friedrich, Extracting strong measurement noise from stochastic time series: Applications to empirical data, *Phys. Rev. E* **81**, 041125 (2010).
- [25] M. B. Weissman, $\frac{1}{f}$ noise and other slow, nonexponential kinetics in condensed matter, *Rev. Mod. Phys.* **60**, 537 (1988).
- [26] G. Bakshi, C. Cao, and Z. Chen, Empirical performance of alternative option pricing models, *J. Finance* **52**, 2003 (1997).
- [27] D. Duffie, J. Pan, and K. Singleton, Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343 (2000).
- [28] T. G. Andersen, L. Benzoni, and J. Lund, An empirical investigation of continuous-time equity return models, *J. Finance* **57**, 1239 (2002).
- [29] S. R. Das, The surprise element: Jumps in interest rates, *J. Econometrics* **106**, 27 (2002).
- [30] M. Johannes, The statistical and economic role of jumps in continuous-time interest rate models, *J. Finance* **59**, 227 (2004).
- [31] Z. Cai and Y. Hong, Some recent developments in nonparametric finance, in *Nonparametric Econometric Methods*, Advances in Econometrics Vol. 25, edited by Q. Li and J. S. Racine (Emerald Group Publishing Limited, Bingley, 2009), pp. 379–432.
- [32] M. Anvari, M. R. R. Tabar, J. Peinke, and K. Lehnertz, Disentangling the stochastic behavior of complex time series, *Sci. Rep.* **6**, 35435 (2016).
- [33] K. Lehnertz, L. Zabawa, and M. R. R. Tabar, Characterizing abrupt transitions in stochastic dynamics, *New J. Phys.* **20**, 113043 (2018).
- [34] C. T. Chudley and R. J. Elliott, Neutron scattering from a liquid on a jump diffusion model, *Proc. Phys. Soc.* **77**, 353 (1961).
- [35] R. C. Merton, Option pricing when underlying stock returns are discontinuous, *J. Financ. Econ.* **3**, 125 (1976).
- [36] P. L. Hall and D. K. Ross, Incoherent neutron scattering functions for random jump diffusion in bounded and infinite media, *Mol. Phys.* **42**, 673 (1981).
- [37] M. T. Giraudo and L. Sacerdote, Jump-diffusion processes as models for neuronal activity, *Biosystems* **40**, 75 (1997).
- [38] R. F. Pawula, Approximation of the linear Boltzmann equation by the Fokker-Planck equation, *Phys. Rev.* **162**, 186 (1967).
- [39] J. Prussek and K. Lehnertz, Stochastic Qualifiers of Epileptic Brain Dynamics, *Phys. Rev. Lett.* **98**, 138103 (2007).
- [40] D. Lamouroux and K. Lehnertz, Kernel-based regression of drift and diffusion coefficients of stochastic processes, *Phys. Lett. A* **373**, 3507 (2009).
- [41] L. R. Gorjão and F. Meirinhos, kramersmoyal: Kramers-Moyal coefficients for stochastic processes *J. Open Source Softw.* (2020), doi: 10.21105/joss.01693.
- [42] The calculations of the KM coefficients are computationally inexpensive. E.g., estimating all 14 KM coefficients as in Fig. 9 takes about 5.5 s on a desktop computer (quad-core 2.20 GHz) for a bidimensional time series of 2×10^6 data points (22 s for 2×10^7). Our approach might be advantageous for field applications that aim at an investigation of interactions between complex systems with poorly understood dynamics.
- [43] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme, Non-gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics, *Nat. Energy* **3**, 119 (2018).
- [44] L. R. Gorjão, M. Anvari, H. Kantz, D. Witthaut, M. Timme, and B. Schäfer, Data-driven model of the power-grid frequency dynamics, *arXiv:1909.08346* (2019).
- [45] M. Anvari, L. R. Gorjão, M. Timme, B. Schäfer, D. Witthaut, and H. Kantz, Stochastic properties of the frequency dynamics in real and synthetic power grids, *arXiv:1909.09110* (2019).

2.3.2 Publication #7

L. Rydin Gorjão, D. Witthaut, and P. G. Lind. *JumpDiff: A Python library for statistical inference of jump-diffusion processes in sets of measurements*, submitted to the Journal of Statistical Software, 2020, Ref. [7].

Status: submitted

JumpDiff: A Python library for statistical inference of jump-diffusion processes in sets of measurements

Leonardo Rydin Gorjão

Forschungszentrum Jülich,
University of Cologne,
Germany

Dirk Witthaut

Forschungszentrum Jülich,
University of Cologne,
Germany

Pedro G. Lind

Dep. Computer Science,
Oslo Metropolitan University,
Norway

Abstract

Models of complex systems based on stochastic processes are ubiquitous across several research fields. However, the complexity of the stochastic contributions lays often beyond pure diffusive processes, such as Brownian motion and random walks, and are better described through so-called jump processes. Jump-diffusion processes incorporate both diffusive contributions as well as stochastic jumps, making them a natural extension of pure diffusive processes towards several applications with sets of measurements. Reliable computation libraries with all necessary tools implemented still lack in software and computer science literature. In this paper we introduce a `python` library, called `JumpDiff`, which includes all necessary functions to assess jump-diffusion processes. One of the strengths of this library is that it includes functions which compute a set of non-parametric estimators of all contributions composing a jump-diffusion process, namely the drift, the diffusion, and the jump strengths. Therefore, having a set of measurements from jump-diffusion process in nature, `JumpDiff` library is able to retrieve the evolution equation producing data series statistically equivalent to the series of measurements. Moreover, the library is also able to test if stochastic jump contributions are presents in the dynamics underlying a set of measurements. The back-end calculations are based in second-order corrections of the condition moments expressed from the series of Kramers–Moyal coefficients. Additionally, we introduce a simple iterative methods for deriving higher-order corrections of any Kramers–Moyal coefficient.

Keywords: Stochastic differential equations, jump-diffusion processes, Kramers–Moyal expansion, Kramers–Moyal coefficients, Python.

1. Introduction

In order to understand processes either riddled or driven by noise, stochastic models can be employed. At the core of several stochastic processes in natural and social sciences lies the Langevin equation, which yields a direct connection to invariants in the process, such as energy and noise contributions (Risken 1984). Because of its fundamental importance, considerably effort has been made to develop methods to estimate non-parametrically from measured data the parameters of stochastic processes, to implement them as numerical routines, as well as extend them both to “processes through scales” (Friedrich and Peinke 1997) and to different complex real systems such as turbulence, wind energy, and brain signal (Friedrich, Peinke, Sahimi, and Tabar 2011). Moreover, complementing those equations with proper numerical schemes, it became possible to apply them to general situations, namely when the set of measurements are subjected to measurement and experimental noise. Such approaches have been already made, for the simplest forms of additive and experimental noise (Böttcher, Peinke, Kleinhans, Friedrich, Lind, and Haase 2006) as well as more general situations, e.g. when measurement noise is time-correlated (Lehle 2011) or when an arbitrary number of stochastic variables and experimental noise sources are coupled with each other simultaneously (Lehle 2013; Scholz, Raischel, Lopes, Lehle, Wächter, Peinke, and Lind 2017).

In recent years, with the goal of going beyond the narrow scope of Gaussian processes, extensions of classical stochastic processes have included jumps, both with the practical aim of better describe the natural world, or by themselves for the riddling mathematical hardship and interesting results. As a direct extension of the classical Langevin equation, jump-diffusion processes came into focus (Duffie, Pan, and Singleton 2000; Johannes 2004; Benth, Di Nunno, and Khedher 2011; Tabar 2019; Rydin Gorjão, Heysel, Lehnertz, and Tabar 2019). This family of processes embodies both a conventional Gaussian (diffusion) contribution as well as a Poissonian (jump) contributions and become particularly hard to tackle, requiring a revised interpretation under the classic Kramers–Moyal expansion. This hardship is given by the presence of jumps and required a revised interpretation under the classic Kramers–Moyal expansion, particularly when the goal is to extract separately Gaussian and jump contributions from a data set. By upgrading purely diffusive assumptions to more general ones, incorporating the possibility of jump-diffusion processes, some of the applications mentioned above could be brought to a new level of knowledge and forecasting capabilities. For example, jump-diffusion processes have found application in the finances (Duffie *et al.* 2000; Andersen, Benzoni, and Lund 2002; Johannes 2004), early-warning signal identification (Dakos, Carpenter, Brock, Ellison, Guttal, Ives, Kéfi, Livina, Seekell, van Nes, and Scheffer 2012), soil moisture dynamics (Daly and Porporato 2006), solar radiation and EEG recordings (Anvari, Lohmann, Wächter, Milan, Lorenz, Heinemann, Tabar, and Peinke 2016a; Anvari, Tabar, Peinke, and Lehnertz 2016b; Lehnertz, Zabawa, and Tabar 2018), neural activity (Giraudo and Sacerdote 1997), amongst others.

In this paper we address two questions: first, we introduce a set of higher-order corrections to the description of particularly simple jump-diffusion processes via the Kramers–Moyal expansion, which links the realm of employing a partial differential representation to stochastic differential representation of the stochastic process. Second, we implement this in a Python library, called JumpDiff, for direct research application. The code is written in Python3 (Van Rossum and Drake 2009). Using this library enables one to extract the parameters of the system in a non-parametrical way, and, in particular, to distinguish between

pure diffusions and jump-diffusions. Our library is suited for the family of Poissonian jump processes, and represents a step forward towards more complex stochastic processes, namely Lévy-like processes (Siegert and Friedrich 2001; Lubashevsky, Friedrich, and Heuer 2009; Zaburdaev, Denisov, and Klafter 2015; Zan, Xu, Kurths, Chechkin, and Metzler 2020).

2. Evolution and data-based inference of jump-diffusion processes

2.1. Theoretical aspects behind jump-diffusion processes

In this section we consider the stochastic evolution of a time-continuous Markov process, $X(t) \in \mathbb{R}$, that is governed by three independent contributions: one drift strength, one diffusive strength, and one Poissonian (jump) strength. The evolution equation of such a variable reads:

$$dX(t) = a(x, t) dt + b(x, t) dW(t) + \xi dJ(t), \quad (1)$$

where $a(x, t)$ is the drift strength, $b(x, t)$ is the diffusion or volatility, $W(t)$ is a Wiener process, and $J(t)$ is a time-homogeneous Poisson jump process with rate $\lambda(x, t)$ and an amplitude ξ , which is normally distributed as $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$.

Jump-diffusion processes governed by Eq. (1) are a generalisation of diffusion processes, since for the latter $\sigma_\xi = 0$ (similarly $\lambda(x, t) = 0$) always. Below, we make use of the so-called Kramers–Moyal expansion to connect the orders of the expansion with each parameter of the jump-diffusion process, and thus enabling their estimation. Notice that the problem of deriving an equation similar to Eq. (1) directly from data sets has been scarcely addressed until very recently (Friedrich *et al.* 2011; Anvari *et al.* 2016b). However, while numerical implementations enabling to model data series with drift and diffusion strengths is already available, cf. (Rinn, Lind, Wächter, and Peinke 2016; Rydin Gorjão and Meirinhos 2019), its extension to incorporate Poissonian contributions is here provided.

In order to link Eq. (1) to the evolution of the probability $p(x, t + \tau | x', t)$, we use the so-called Kramers–Moyal equation

$$\frac{\partial}{\partial \tau} p(x, t + \tau | x', t) = \mathcal{L}_{KM} p(x, t + \tau | x', t), \quad (2)$$

where \mathcal{L}_{KM} denotes the Kramers–Moyal operator defined as (Risken 1984)

$$\mathcal{L}_{KM} = \sum_{m=1}^{\infty} \left(-\frac{\partial}{\partial x} \right)^m D_m(x). \quad (3)$$

Functions $D_m(x)$ are called Kramers–Moyal coefficients relate to the m -order conditional moment $M_m(x, \tau)$, given by

$$M_m(x, t, \tau) = \int_{-\infty}^{\infty} (x' - x)^m p(x', t + \tau | x, t) dx'. \quad (4)$$

For simplicity we drop the t -dependency focusing on stationary processes. The Kramers–Moyal coefficients $D_m(x)$ are thus defined for any integer m as

$$D_m(x) = \frac{1}{m!} \lim_{\tau \rightarrow 0} \frac{M_m(x, \tau)}{\tau}, \quad (5)$$

The jump-diffusion process defined in Eq. (1) is linked to a particular case of the Kramers–Moyal expansion defined in Eqs. (2) and (3), namely

$$\frac{\partial}{\partial \tau} p(x, t + \tau | x', t) = \left[-a(x, t) \frac{\partial}{\partial x} + \left(b(x, t)^2 + \lambda(x, t) \sigma_\xi \right) \frac{\partial^2}{\partial x^2} + \sum_{2k=4}^{\infty} \sigma_\xi^k \lambda \frac{\partial^{2k}}{\partial x^{2k}} \right] p(x, t + \tau | x', t). \quad (6)$$

For purely diffusive processes, Eq. (2) reduces to the Fokker–Planck–Kolmogorov equation (sometimes denoted solely Fokker–Planck or Smoluchowski equation) (Risken 1984), and the estimates of the two first Kramers–Moyal coefficients are already available (Friedrich *et al.* 2011), even employing higher-order approximations (Gottschall and Peinke 2008). In the case of jump-diffusion processes, higher-order coefficients should be taken into account as they are non-vanishing.

For instance, in jump-diffusion processes (Anvari *et al.* 2016b) one typically needs the Kramers–Moyal coefficients up to sixth order. Up to now, higher-order Kramers–Moyal coefficient were estimated from first-order estimations of the conditional moments. Here, we derive the second-order approximations, which imply a more cumbersome analytical approach to the Kramers–Moyal Eq. (2).

By defining all coefficients $D_m(x)$ one is able to define a model for the conditional probability distribution of a given process, allowing one to (numerically) generate samples of that process. The inverse problem however is not so trivial and implies deriving estimates of the coefficients $D_m(x)$ from sampled series of the process.

2.2. Numerical computation and approximations of the Kramers–Moyal coefficients

The numerical computation of the Kramers–Moyal coefficients is based on the numerical computation of conditional moments $M_m(x, t)$, which can be estimated directly from a set of measurements $X(t)$. Indeed, the instantaneous time rate of the moment of order n for the process $X(t)$, conditioned to a specific value x , is given by

$$M_m(x, \tau) = \langle (X(t + \tau) - X(t))^m | X(t) = x \rangle, \quad (7)$$

with $\langle X(t) \rangle$ denoting the average of $X(t)$, for all measured t . For pure diffusive processes, it is only needed to compute the first and second conditional moment ($m = 1, 2$). For jump-diffusion processes the spectrum of needed moments extends beyond the first two conditional moments up to the sixth-order moment, noting here that all moments exist (Anvari *et al.* 2016b).

The conditional moments can be expressed as sums of products of Kramers–Moyal coefficients, derived from the formal solution of Eq. (2), namely

$$p(x, t + \tau | x', t) = \exp(\tau \mathcal{L}_{\text{KM}}) \delta(x - x') = \sum_{k=0}^{\infty} \frac{(\tau \mathcal{L}_{\text{KM}})^k}{k!} \delta(x - x'). \quad (8)$$

Depending on the number of terms used from the sum in Eq. (8) one obtains different orders of approximation of the Kramers–Moyal operator. Here we will consider first- and second-order approximations.

First-order approximation

The first-order approximation is given by

$$\exp(\tau \mathcal{L}_{\text{KM}}) \sim 1 + \tau \mathcal{L}_{\text{KM}}, \quad (9)$$

where expressing the Kramers–Moyal coefficients from Eq. (5), via the moments in Eq. (4) yields

$$D_m(x) = \frac{1}{m!} \lim_{\tau \rightarrow 0} \frac{M_m(x, \tau)}{\tau}. \quad (10)$$

The full derivation is given in Appendix A.

Using this approximation the system mapping between six conditional moments $M_n(x, t)$ and the parameters defining Eq. (1) are given by (Anvari *et al.* 2016b):

$$a(x, t) = D_1(x, t), \quad (11a)$$

$$b^2(x, t) = D_2(x, t) - \lambda(x, t)\sigma_\xi^2, \quad (11b)$$

$$\sigma_\xi^2 = \frac{D_6(x, t)}{5D_4(x, t)}. \quad (11c)$$

$$\lambda(x, t) = \frac{D_4(x, t)}{3\sigma_\xi^4}, \quad (11d)$$

Notice the relation of the parameters of a jump-diffusion process in Eq. (1) to the conditional moments in Eq. (7) is given with all generality for higher-order terms by $D_{2m} = (2m!) \sigma_\xi^m \lambda$, for $m \geq 3$ (Anvari *et al.* 2016b).

Second-order approximation

Higher-order approximations of the conditional moment are especially relevant when handling low-sampled data. The second-order approximation of the conditional moments takes in one additional term from the sum in Eq. (8), namely

$$\exp(\tau \mathcal{L}_{\text{KM}}) \sim 1 + \tau \mathcal{L}_{\text{KM}} + \frac{\tau^2}{2} \mathcal{L}_{\text{KM}} \mathcal{L}_{\text{KM}}, \quad (12)$$

where naturally the first-order terms $\sim \tau$ are present at the first-order of the approximation. The derivation, found in full detail in Appendix A, involves a set of relations between the moments $M_m(x, \tau)$ of a given order m and the Kramers–Moyal coefficients $D_n(x)$ of all orders $0 < n \leq m$. By inverting these relations with respect to the Kramers–Moyal coefficients yields

$$D_m(x) = \frac{1}{m!} \lim_{\tau \rightarrow 0} \frac{F_m(x, \tau)}{\tau}, \quad (13)$$

where $F_m(x, \tau)$ denotes the second-order approximation, in comparison with the first-order approximation given above in Eq. (10). The second-order approximations $F_m(x, \tau)$ are given

FUNCTION	PARAMETERS	OUTPUTS	INTERNAL LIBRARIES	EXTERNAL LIBRARIES
jdprocess()	time , delta_t , drift a , diffusion b , jump amplitude xi , jump rate lamb	timeseries X	None	numpy
moments()	timeseries , bins bins , order power , time lag lag , correction	space edges , moments	binning , kernels	numpy , scipy
jump_amplitude()	moments	estimator xi_est	None	numpy
jump_rate()	moments , jump amplitude xi_est	estimator lamb_est	None	numpy
Qratio()	time lag lag , timeseries	time lag lag , ratio ratio	moments	numpy
corrections()	moments m , order power	corrected moments	None	numpy
M_formula()	order power	symbolic term	None	sympy
F_formula()	order power	symbolic term	None	sympy

Table 1: Primary functions (top) and helping functions (bottom) implemented in the JumpDiff library.

by (dependencies removed for clarity)

$$F_1 = M_1, \quad (14a)$$

$$F_2 = M_2 - M_1^2, \quad (14b)$$

$$F_3 = M_3 - 3M_1M_2 + 3M_1^3, \quad (14c)$$

$$F_4 = M_4 - 4M_1M_3 + 18M_1^2M_2 - 3M_2^2 - 15M_1^4, \quad (14d)$$

$$F_5 = M_5 - 5M_1M_4 + 30M_1^2M_3 - 150M_1^3M_2 + 45M_1M_2^2 - 10M_2M_3 + 105M_1^5, \quad (14e)$$

$$F_6 = M_6 - 6M_1M_5 + 45M_1^2M_4 - 300M_1^3M_3 + 1575M_1^4M_2 - 675M_1^2M_2^2 \\ + 180M_1M_2M_3 + 45M_2^3 - 15M_2M_4 - 10M_3^2 - 945M_1^6. \quad (14f)$$

Where naturally the first term on each right-hand side is the first-order approximation. This second-order approximation neglects terms including derivatives of the Kramers–Moyal coefficients, which enables one to express the n -th Kramers–Moyal coefficient as a function of conditional moments up to order $n-1$. In this way, we provide a general formula for improving the estimates of Kramers–Moyal coefficient, generally taken as linear approximations of the corresponding conditional moment.

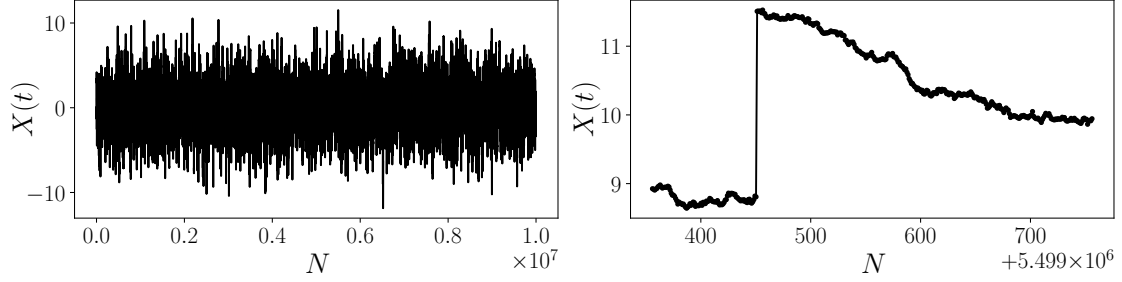


Figure 1: Illustration of a jump-diffusion process using function `jdprocess` which implements Eq. (15). (Left) the full extent of Eq. (15) with $N = 10^7$. (Right) an exemplary jump in the integrated process. Here $\theta = 0.5$, $\sigma = 0.75$, $\sigma_\xi^2 = 1.5$, and $\lambda = 1.75$.

3. Implementation of the functions in library `JumpDiff`

In this section, we describe how we implement all functions for deriving a jump-diffusion equation for stochastic processes, using the approximations above. All the functions are implemented in the `python` library `JumpDiff` and are listed in Table (1).

Codes using the library `JumpDiff` must import the package, eventually attributing another name, e.g.

```
import JumpDiff as jd
```

Notice that library `JumpDiff` requires libraries `numpy` (van der Walt, Colbert, and Varoquaux 2011), `scipy` (Virtanen, Gommers, Oliphant, Haberland, Reddy, Cournapeau, Burovski, Peterson, Weckesser, Bright, van der Walt, Brett, Wilson, Jarrod Millman, Mayorov, Nelson, Jones, Kern, Larson, Carey, Polat, Feng, Moore, Vand erPlas, Laxalde, Perktold, Cimrman, Henriksen, Quintero, Harris, Archibald, Ribeiro, Pedregosa, van Mulbregt, and Contributors 2020), and `sympy` (Meurer, Smith, Paprocki, Čertík, Kirpichev, Rocklin, Kumar, Ivanov, Moore, Singh, Rathnayake, Vig, Granger, Muller, Bonazzi, Gupta, Vats, Johansson, Pedregosa, Curry, Terrel, Roučka, Saboo, Fernando, Kulal, Cimrman, and Scopatz 2017).

3.1. From equation to data: generating sample trajectories of jump-diffusion processes

The package `JumpDiff` contains the function `jdprocess()` capable of generating sample trajectories of a jump-diffusion process. We will generate a single trajectory of the process, and subsequently employ the non-parametric estimators in `JumpDiff` to retrieve the parameters of the jump-diffusion process generated, described in the following subsections.

To compare the first-order and second-order approximations derived above we first numerically generate a single trajectory of an Ornstein–Uhlenbeck process with Poissonian jumps, namely

$$dX = -\theta x dt + \sigma dW(t) + \xi dJ(t). \quad (15)$$

Function `jdprocess()` is used for integrating a jump-diffusion process. As an example, the numerical integration of Eq. (15), with a number of points $N = 1 \times 10^7$ ($t = 10^4$) and a time-step of $\Delta t = 0.001$, can be implemented as follows:

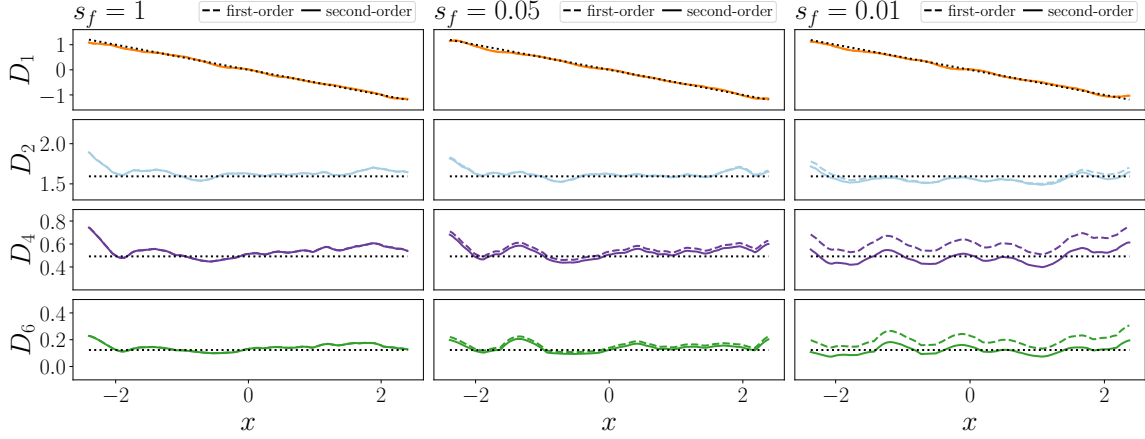


Figure 2: Kramers–Moyal coefficients $D_n(x)$ computed with first-order approximations (dashed lines) and second-order approximations (solid lines). The data set used for computing Kramers–Moyal coefficients was generated by integrating Eq. (15) with the same parameter values as in Fig. 1. Three sampling rates are consider: (left) every integrated value, $s_f = 1$, (middle) $s_f = 0.05$, (right) $s_f = 0.01$. The dotted line indicates the theoretical result.

```
# integration time and time sampling
time = 10000
delta_t = 0.001
# define the drift function a(x)
def a(x):
    return -0.5*x
# define the diffusion function b(x)
def b(x):
    return 0.75
# define jump height and rate
xi = 1.5
lamb = 1.75
# generate the jump-diffusion process
X = jd.jdprocess(time, delta_t, a, b, xi, lamb)
```

Fig. 1 shows the trajectory generated with this code.

While in this example an Ornstein–Uhlenbeck process, i.e., linear drift strength and constant diffusion strength, is chosen with a Poissonian jump strength, other higher polynomial functions can be chosen for the different contributions by adjusting the drift and diffusion functions.

3.2. From data to the jump-diffusion equation: the inverse problem

Having described how to generate series of values from a jump-diffusion equation, using the function `jdprocess()`, we now consider the inverse problem: starting from that series of values, derive the parameters of the jump-diffusion equation. To that end, we extract the Kramers–Moyal coefficients $D_m(x)$.

Fig. 2 shows the derived results for the inverse problem, using first-order (dashed lines) and second-order corrections (solid lines). The theoretical values are indicated by the dotted lines. Furthermore, by down-sampling the process one studies the impact of second-order corrections in a low-sampled data. We consider different sample rates, namely $s_f = 1, 0.05$, and 0.01 . As can be seen in Fig. 2 second-order corrections improve the estimate of all Kramers–Moyal coefficients $D_m(x)$, $m \geq 2$, especially in the cases with lowest sampling rates.

The Kramers–Moyal coefficients are estimated using the function **moments()**:

```
# Choose bandwidth of kernel
bw = 0.35
# extract the Kramers--Moyal and space without second-order corrections
edge, simple_mom = jd.moments(timeseries = X, bw = bw, correction = False)
# and with second-order terms
edge, mom = jd.moments(timeseries = X, bw = bw, correction = True)
```

The function **moments()** performs a kernel-density estimation employing a Nadaraya–Watson estimator (Nadaraya 1964; Watson 1964) of the different orders of the moments. By default it employs an Epanechnikov kernel and uses a convolution method to perform the kernel-density estimation of the moments. Four other kernels are also available, namely Gaussian, uniform, triangular, and quartic kernels. Moreover, the function **moments()** includes the input parameter **lag**, which is a time-lag that enables one to estimate the conditional moments at different time-lags τ . If left unspecified, it assumes the shortest increment of the timeseries, i.e., the timeseries sampling rate $\tau = 1/s_f$. This is especially suited for evaluating the conditional moments in the limiting case of $\tau \rightarrow 0$, since numerical accuracy is bounded by the sampling rate s_f . This evaluation is done by plotting a few time-lags, e.g. $\tau = 1/s_f, \dots, 10/s_f$, and extrapolate the limit $\tau \rightarrow 0$ (Böttcher *et al.* 2006; Lind, Haase, Böttcher, Peinke, Kleinhans, and Friedrich 2010).

3.3. Extracting all parameters for a single trajectory

In Fig. 2 we can see the estimation of the Kramers–Moyal coefficients, with first- and second-order corrections. The results reproduce well the theoretical values (dotted lines), which allows us to further extract the parameters of our process via the Eqs. (11). Retrieving the drift and diffusion functions, $a(x, t)$ and $b(x, t)$ respectively, are known problems, which imply studying the first and second Kramers–Moyal coefficients.

From Eqs. (11) one recovers also the jump amplitude σ_ξ^2 and the jump rate λ of the time series, by considering higher-order conditional moments. In **JumpDiff** library, functions **jump_rate()** and **jump_amplitude()** implement Eqs. (11d) and (11c), respectively. The examples plotted in Fig. 2 were generated with the following implementation:

```
# Take the timeseries X to obtain xi
sigma_xi_est = jd.jump_amplitude(timeseries = X)
# And to obtain lamb
lamb_est = jd.jump_rate(timeseries = X)
```

In Fig. 3 we numerically integrate Eq. (15), as before, and employ the estimators of the jump amplitude and jump rate, for a timeseries with increasing number of data points N . The

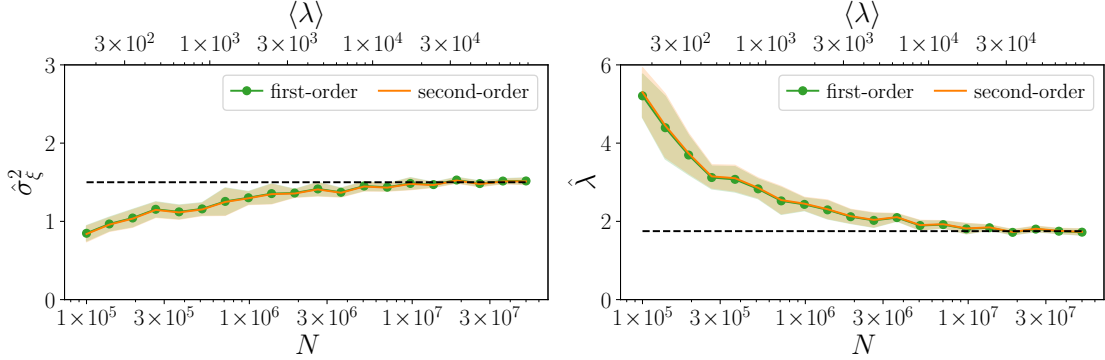


Figure 3: Estimation of the jump amplitude $\hat{\sigma}_\xi^2$ (left) and the jump rate $\hat{\lambda}$ (right) from 20 numerically simulated jump-diffusion processes with increasing time lengths N . The data was generated by integrating Eq. (15), with the $\sigma_\xi = 1.5$ and $\lambda = 1.75$ (dashed lines), and each term was estimated using the functions `jump_amplitude()` and `jump_rate()`. The estimates are shown as a function of the number of points $N \in [1 \times 10^5, 5 \times 10^7]$, always with a time-step $\Delta t = 0.001$. Top axis denotes the average number of jumps $\langle \lambda \rangle$ in the respective numerically integrated time series. Each point is an average over 10 iterations. Standard deviations depicted in the shaded areas.

	$a(x)$	$b(x)$	σ_ξ^2	λ
Theoretical	$-0.5x$	0.75	1.5	1.75
Estimated	$-0.496x$	0.760	1.524	1.802

Table 2: Parameter estimation for the process Eq. (15) in Fig. 1, generated with `jdprocess()`, with indicated parameters.

estimators converge to the theoretical values (dashed line) with increasing accuracy with the average number on jumps $\langle \lambda \rangle$ in the timeseries. The comparison between theoretical values and estimated parameters is given in Tab. 2. For a general method to recover all Kramers–Moyal coefficients, for stochastic processes of any dimension, see Ref. (Rydin Gorjão and Meirinhos 2019).

3.4. Evaluating if the process is purely a diffusion or a jump-diffusion

The presence of the jump contribution in Eq. (1) is the fundamental addition to a general diffusion equation. However, assuming such *Ansatz* for purely diffusive process may lead to spurious jump terms. To avoid this, one fundamental question to ask is whether we are in the presence of a pure diffusion or a jump-diffusion process, in order to choose *ab initio* the proper *Ansatz*.

Jump-diffusion processes—and in general jump processes—display higher-order conditional moments. In Ref. (Lehnertz *et al.* 2018) the authors introduced a simple criteria to distinguish between pure diffusive and jump-diffusions, which is based in the fourth- and sixth-order

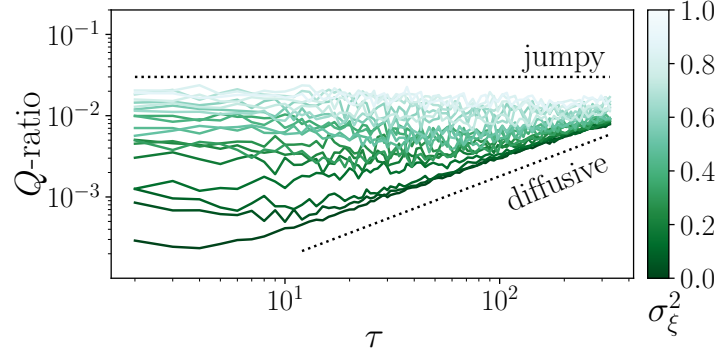


Figure 4: Illustration of the Q -ratio, defined in Eq. (16), for the process in Eq. (1) with $a(x) = -x$ and $b(x) = 1$ and varying jump amplitudes $\sigma_\xi^2 \in [0, 1]$ with a fixed jump rate $\lambda = 0.1$. The equation was numerically integrated with $t = 1000$ and timestep 0.001. Each line is an average of 20 iterations.

moments. Defining the Q -ratio as

$$Q(x, \tau) = \frac{M_6(x, \tau)}{5M_4(x, \tau)}, \quad (16)$$

if the process is purely diffusive $Q(x, \tau) = \tau(b(x))^2$ (linear function with τ), whereas if the process has a jump term, $Q(x, \tau) = \sigma_\xi^2$ (constant). This criterion can be employed directly for any time series and is implemented as the function `Qratio` in library `JumpDiff`.

```
# Take a sequence of integers lag
lag = logspace(0, 4, 25, dtype=int)
# Recover the Q-ratio of the timeseries X
lag, Q = jd.Qratio(lag, X)
# plot in a log-log scale
plt.loglog(lag, Q)
```

Fig. 4 illustrates the implementation of the function `Qratio()`, plotting it for several time-lags τ in a log-log plot. If there is a linear dependence on τ the time series $X(t)$ is a pure diffusive process, whereas if the plot is approximately flat, showing a constant Q -ratio the time series should have a jump term. Notice that increasing the jump rate from $\xi = 0$ (pure diffusion) to $\xi = b(x) = 1$ (same amplitude as diffusion strength), the Q -ratio changes from linearly depending on τ to a constant.

3.5. Kramers–Moyal coefficients of arbitrary order

The Kramers–Moyal coefficients given in Eq. (13) can be given for arbitrary order n , by implementing the following general expression

$$D_n(x) = \frac{1}{\tau(n!)} \left[\hat{B}_{n,1}(M_1(x, \tau), M_2(x, \tau), \dots, M_n(x, \tau)) - \frac{\tau}{2} \hat{B}_{n,2}(M_1(x, \tau), M_2(x, \tau), \dots, M_{n-1}(x, \tau)) \right],$$

where $\hat{B}(\cdot)$ refer to the ordinary Bell polynomials. The first term on the right-hand side account for the first-order corrections, while the second term accounts for second-order corrections. The derivation of this expression is given in Appendix B.

This formula and its reciprocal, which accounts for the relation of the conditional moments $M_m(x, \tau)$ with the Kramers–Moyal coefficients $D_m(x)$, are implemented in functions **F_formula()** and **M_formula()**, respectively, where python’s symbolic language `sympy` is used.

4. Discussion and conclusions

Jump-diffusion are stochastic models able to describe processes riddled with jumps. Alongside with regular diffusion processes, their elegance lies in the possibility of estimating non-parametrically the parameters via the Kramers–Moyal expansion. Access to higher-order moments of the Kramers–Moyal equation is computationally feasible, and thus permits a one-to-one correspondence between model parameters and the estimators (cf. Eq. (11)). These on the other hand are hampered by an assumption already existent at the level of the model: a scarcity of total jumps. Where the diffusive strengths are ubiquitous, the jumps take place sparsely in time. To this effect are corrections implemented here crucial to a correct retrieval of the parameters of the model.

The application of jump-diffusion models as descriptive of processes beyond diffusion is applicable across different research fields, as evidenced by the vast publications mentioned in the introduction. The elegance of the process lies not only on the general applicability, but also the ease of applying non-parametric parameter estimation. In this sense, the presented higher-order corrections are of substantial relevance, especially taking into account jumps occur sparsely. Here we present a two-fold project, developing second-order corrections of the Kramers–Moyal expansion of jump-diffusion processes and designing a easy-to-employ python library.

The limitations and finite scope of the library here described motivate further development, namely towards more general jumps types, having amplitudes and rates which are time (and space) dependent. Moreover, beyond Poissonian jumps, new libraries can be developed, for instance to approach Lévy-like processes.

Notice here that the problem is invertible in one dimension, i.e., knowing the conditional moments allows one to know the parameters of the process, but it does not generalise for higher-dimensions, as is already the case for diffusions. Two-dimensional jump-diffusion processes have been addressed in Ref. (Rydin Gorjão *et al.* 2019).

5. Acknowledgements

The authors thank Sílvia M. D. Queirós for useful discussions. L. R. G. and D. W. gratefully acknowledges support by the Helmholtz Association, via the joint initiative *Energy System 2050 – A Contribution of the Research Field Energy*, the grant No. VH-NG-1025, the associative *Uncertainty Quantification – From Data to Reliable Knowledge (UQ)* with grant no. ZT-I-0029. L. R. G gratefully acknowledges the scholarship funding from *E.ON Stipendienfonds* and the *STORM – Stochastics for Time-Space Risk Models* project of the Research Council of Norway (RCN) No. 274410. This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

References

- Andersen TG, Benzoni L, Lund J (2002). “An Empirical Investigation of Continuous-Time Equity Return Models.” *The Journal of Finance*, **57**(3), 1239–1284. doi:10.1111/1540-6261.00460.
- Anvari M, Lohmann G, Wächter M, Milan P, Lorenz E, Heinemann D, Tabar MRR, Peinke J (2016a). “Short term fluctuations of wind and solar power systems.” *New Journal of Physics*, **18**, 063027. doi:10.1088/1367-2630/18/6/063027.
- Anvari M, Tabar MRR, Peinke J, Lehnertz K (2016b). “Disentangling the stochastic behavior of complex time series.” *Scientific Reports*, **6**, 35435. doi:10.1038/srep35435.
- Benth F, Di Nunno G, Khedher A (2011). “Robustness of option prices and their deltas in markets modeled by jump-diffusions.” *Communications on Stochastic Analysis*, **5**(2), 285–307. doi:10.31390/cosa.5.2.03.
- Böttcher F, Peinke J, Kleinhans D, Friedrich R, Lind PG, Haase M (2006). “Reconstruction of complex dynamical systems affected by strong measurement noise.” *Physical Review Letters*, **97**, 090603. doi:10.1103/PhysRevLett.97.090603.
- Dakos V, Carpenter SR, Brock WA, Ellison AM, Guttal V, Ives AR, Kéfi S, Livina V, Seekell DA, van Nes EH, Scheffer M (2012). “Methods for Detecting Early Warnings of Critical Transitions in Time Series Illustrated Using Simulated Ecological Data.” *PLoS ONE*, **7**(7), 1–20. doi:10.1371/journal.pone.0041010.
- Daly E, Porporato A (2006). “Probabilistic dynamics of some jump-diffusion systems.” *Physical Review E*, **73**, 026108. doi:10.1103/PhysRevE.73.026108.
- Duffie D, Pan J, Singleton K (2000). “Transform Analysis and Asset Pricing for Affine Jump-diffusions.” *Econometrica*, **68**(6), 1343–1376. doi:10.1111/1468-0262.00164.
- Friedrich R, Peinke J (1997). “Description of a Turbulent Cascade by a Fokker–Planck Equation.” *Physical Review Letters*, **78**, 863–866. doi:10.1103/PhysRevLett.78.863.
- Friedrich R, Peinke J, Sahimi M, Tabar MRR (2011). “Approaching complexity by stochastic methods: From biological systems to turbulence.” *Physics Reports*, **506**(5), 87–162. doi:10.1016/j.physrep.2011.05.003.
- Giraud MT, Sacerdote L (1997). “Jump-diffusion processes as models for neuronal activity.” *Biosystems*, **40**(1), 75–82. doi:10.1016/0303-2647(96)01632-2.
- Gottschall J, Peinke J (2008). “On the definition and handling of different drift and diffusion estimates.” *New Journal of Physics*, **10**, 083034. doi:10.1088/1367-2630/10/8/083034.
- Johannes M (2004). “The Statistical and Economic Role of Jumps in Continuous-Time Interest Rate Models.” *The Journal of Finance*, **59**(1), 227–260. doi:10.1111/j.1540-6321.2004.00632.x.
- Lehle B (2011). “Analysis of stochastic time series in the presence of strong measurement noise.” *Physical Review E*, **83**, 021113. doi:10.1103/PhysRevE.83.021113.

- Lehle B (2013). “Stochastic Time Series with Strong, Correlated Measurement Noise: Markov Analysis in N Dimensions.” *Journal of Statistical Physics*, **152**, 1145. doi: [10.1007/s10955-013-0803-z](https://doi.org/10.1007/s10955-013-0803-z).
- Lehnertz K, Zabawa L, Tabar MRR (2018). “Characterizing abrupt transitions in stochastic dynamics.” *New Journal of Physics*, **20**(11), 113043. doi: [10.1088/1367-2630/aaf0d7](https://doi.org/10.1088/1367-2630/aaf0d7).
- Lind PG, Haase M, Böttcher F, Peinke J, Kleinhans D, Friedrich R (2010). “Extracting strong measurement noise from stochastic time series: Applications to empirical data.” *Phys. Rev. E*, **81**, 041125. doi: [10.1103/PhysRevE.81.041125](https://doi.org/10.1103/PhysRevE.81.041125).
- Lubashevsky I, Friedrich R, Heuer A (2009). “Continuous-time multidimensional Markovian description of Lévy walks.” *Physical Review E*, **80**(2), 031148. doi: [10.1103/PhysRevE.80.031148](https://doi.org/10.1103/PhysRevE.80.031148).
- Meurer A, Smith CP, Paprocki M, Čertík O, Kirpichev SB, Rocklin M, Kumar A, Ivanov S, Moore JK, Singh S, Rathnayake T, Vig S, Granger BE, Muller RP, Bonazzi F, Gupta H, Vats S, Johansson F, Pedregosa F, Curry MJ, Terrel AR, Roučka u, Saboo A, Fernando I, Kulal S, Cimrman R, Scopatz A (2017). “SymPy: symbolic computing in Python.” *PeerJ Computer Science*, **3**, e103. ISSN 2376-5992. doi: [10.7717/peerj-cs.103](https://doi.org/10.7717/peerj-cs.103).
- Nadaraya EA (1964). “On Estimating Regression.” *Theory of Probability & Its Applications*, **9**(1), 141–142. doi: [10.1137/1109020](https://doi.org/10.1137/1109020).
- Rinn P, Lind PG, Wächter M, Peinke J (2016). “The Langevin Approach: An R Package for Modeling Markov Processes.” *Journal of Open Research Software*, **4**, e34. doi: [10.5334/jors.123](https://doi.org/10.5334/jors.123).
- Risken H (1984). *The Fokker–Planck equation*. Springer, Berlin, Heidelberg. ISBN 978-3-642-61544-3. doi: [10.1007/978-3-642-61544-3](https://doi.org/10.1007/978-3-642-61544-3).
- Rydin Gorjão L, Heysel J, Lehnertz K, Tabar MRR (2019). “Analysis and data-driven reconstruction of bivariate jump-diffusion processes.” *Physical Review E*, **100**, 062127. doi: [10.1103/PhysRevE.100.062127](https://doi.org/10.1103/PhysRevE.100.062127).
- Rydin Gorjão L, Meirinhos F (2019). “kramersmoyal: Kramers–Moyal coefficients for stochastic processes.” *Journal of Open Source Software*, **4**(44). doi: [10.21105/joss.01693](https://doi.org/10.21105/joss.01693).
- Scholz T, Raischel F, Lopes V, Lehle B, Wächter M, Peinke J, Lind PG (2017). “Parameter-free resolution of the superposition of stochastic signals.” *Physics Letters A*, **381**, 194–206. doi: [10.1016/j.physleta.2016.09.057](https://doi.org/10.1016/j.physleta.2016.09.057).
- Siebert S, Friedrich R (2001). “Modeling of nonlinear Lévy processes by data analysis.” *Physical Review E*, **64**(2), 041107. doi: [10.1103/PhysRevE.64.041107](https://doi.org/10.1103/PhysRevE.64.041107).
- Tabar MRR (2019). *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems*. Springer International Publishing. ISBN 978-3-030-18471-1. doi: [10.1007/978-3-030-18472-8](https://doi.org/10.1007/978-3-030-18472-8).
- van der Walt S, Colbert SC, Varoquaux G (2011). “The NumPy Array: A Structure for Efficient Numerical Computation.” *Computing in Science Engineering*, **13**(2), 22–30. doi: [10.1109/MCSE.2011.37](https://doi.org/10.1109/MCSE.2011.37).

- Van Rossum G, Drake FL (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA. ISBN 1441412697.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Jarrod Millman K, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey C, Polat İ, Feng Y, Moore EW, Vand erPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Contributors S (2020). “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods*, **17**, 261–272. doi:10.1038/s41592-019-0686-2.
- Watson GS (1964). “Smooth Regression Analysis.” *Sankhyā: The Indian Journal of Statistics, Series A*, **26**(4), 359–372. ISSN 0581572X. URL <http://www.jstor.org/stable/25049340>.
- Zaburdaev V, Denisov S, Klafter J (2015). “Lévy walks.” *Review of Modern Physics*, **87**(2), 483. doi:10.1103/RevModPhys.87.483.
- Zan W, Xu Y, Kurths J, Chechkin AV, Metzler R (2020). “Stochastic dynamics driven by combined Lévy-Gaussian noise: fractional Fokker–Planck–Kolmogorov equation and solution.” *Journal of Physics A: Mathematical and Theoretical*, **53**(38), 385001. doi:10.1088/1751-8121/aba654.

A. Second-order corrections of Kramers–Moyal coefficients

For the first-order approximation of conditional moments M_n , we substitute in Eq. (4) the first-order approximation of the exponential operator, namely

$$\exp(\tau \mathcal{L}_{\text{KM}}) \sim 1 + \tau \mathcal{L}_{\text{KM}},$$

yielding

$$\begin{aligned} M_n(x', \tau) &\sim \int_{-\infty}^{\infty} (x - x')^n (1 + \tau \mathcal{L}_{\text{KM}}) \delta(x - x') dx \\ &= \int_{-\infty}^{\infty} (x - x')^n \delta(x - x') dx + \tau \int_{-\infty}^{\infty} (x - x')^n \left[\sum_{m=1}^{\infty} \left(-\frac{d}{dx} \right)^m D_m(x) \right] \delta(x - x') dx \\ &= 0 + \tau \sum_{m=1}^{\infty} (-1)^m D_m(x') \int_{-\infty}^{\infty} (x - x')^n \left(\frac{d}{dx} \right)^m \delta(x - x') dx. \end{aligned} \quad (17)$$

The last integral is given by:

$$\begin{aligned} I_1 &= \int_{-\infty}^{\infty} (x - x')^n \left(\frac{d}{dx} \right)^m \delta(x - x') dx \\ &= \left[(x - x')^n \left(\frac{d}{dx} \right)^{m-1} \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} n(x - x')^{n-1} \left(\frac{d}{dx} \right)^{m-1} \delta(x - x') dx \\ &= 0 - \int_{-\infty}^{\infty} n(x - x')^{n-1} \left(\frac{d}{dx} \right)^{m-1} \delta(x - x') dx \\ &= \begin{cases} (-1)^n (n!) \int_{-\infty}^{\infty} \left(\frac{d}{dx} \right)^{m-n} \delta(x - x') dx & \Leftarrow m \geq n \\ (-1)^m \frac{n!}{(n-m-1)!} \int_{-\infty}^{\infty} (x - x')^{n-m-1} \delta(x - x') dx & \Leftarrow m < n \end{cases} \\ &= (-1)^n (n!) \delta_{nm} \end{aligned}$$

yielding

$$M_n(x', \tau) \sim (n!) \tau D_n(x'). \quad (18)$$

This approximation is the one used in Ref. (Anvari *et al.* 2016b).

For the second-order approximation of conditional moments M_n , we substitute in Eq. (4) the second-order approximation of the exponential operator, namely

$$\exp(\tau \mathcal{L}_{\text{KM}}) \sim 1 + \tau \mathcal{L}_{\text{KM}} + \frac{\tau^2}{2} \mathcal{L}_{\text{KM}} \mathcal{L}_{\text{KM}}, \quad (19)$$

yielding

$$\begin{aligned}
M_n(x', \tau) &\sim \int_{-\infty}^{\infty} (x - x')^n \left(1 + \tau \mathcal{L}_{\text{KM}} + \frac{\tau^2}{2} \mathcal{L}_{\text{KM}} \mathcal{L}_{\text{KM}} \right) \delta(x - x') dx \\
&= (n!) \tau D_n(x') + \frac{\tau^2}{2} \int_{-\infty}^{\infty} (x - x')^n \mathcal{L}_{\text{KM}} \mathcal{L}_{\text{KM}} \delta(x - x') dx \\
&= (n!) \tau D_n(x') + \frac{\tau^2}{2} \int_{-\infty}^{\infty} (x - x')^n \left[\sum_{p=1}^{\infty} \left(-\frac{d}{dx} \right)^p D_p(x) \right] \left[\sum_{m=1}^{\infty} \left(-\frac{d}{dx} \right)^m D_m(x) \right] \delta(x - x') dx \\
&= (n!) \tau D_n(x') + \frac{\tau^2}{2} \sum_{p=1}^{\infty} \sum_{m=1}^{\infty} \int_{-\infty}^{\infty} (x - x')^n \left(-\frac{d}{dx} \right)^p D_p(x) \left(-\frac{d}{dx} \right)^m D_m(x) \delta(x - x') dx \\
&= (n!) \tau D_n(x') + \frac{\tau^2}{2} \sum_{p=1}^{\infty} \sum_{m=1}^{\infty} (-1)^{p+m} D_m(x') \int_{-\infty}^{\infty} (x - x')^n \left(\frac{d}{dx} \right)^p D_p(x) \left(\frac{d}{dx} \right)^m \delta(x - x') dx.
\end{aligned}$$

The last integral is derived as follows:

$$\begin{aligned}
I_2 &= \int_{-\infty}^{\infty} (x - x')^n \left(\frac{d}{dx} \right)^p D_p(x) \left(\frac{d}{dx} \right)^m \delta(x - x') dx \\
&= \int_{-\infty}^{\infty} (x - x')^n \sum_{s=0}^p \binom{p}{s} \left[\left(\frac{d}{dx} \right)^{p-s} D_p(x) \right] \left[\left(\frac{d}{dx} \right)^{m+s} \delta(x - x') \right] dx \\
&= \sum_{s=0}^p \binom{p}{s} \int_{-\infty}^{\infty} G(x) \left(\frac{d}{dx} \right)^{m+s} \delta(x - x') dx
\end{aligned} \tag{20}$$

with

$$G(x) = (x - x')^n \left(\frac{d}{dx} \right)^{p-s} D_p(x).$$

The last member of Eq. (20) is computed in a similar way as integral I_1 , yielding

$$\begin{aligned}
I_2 &= \sum_{s=0}^p \binom{p}{s} (-1)^{m+s} \left[\frac{d^{m+s}}{dx^{m+s}} F(x) \right]_{x=x'} \\
&= \sum_{s=0}^p \binom{p}{s} (-1)^{m+s} \sum_{q=0}^{m+s} \binom{m+s}{q} \left[\frac{d^{m+s-q}}{dx^{m+s-q}} (x - x')^n \right]_{x=x'} \left[\frac{d^{p-s+q}}{dx^{p-s+q}} D_p(x) \right]_{x=x'}. \tag{21}
\end{aligned}$$

If $m + s - q > n$, the derivative of $(x - x')^n$ is zero; if $m + s - q < n$, the derivative of $(x - x')^n$ is proportional to $(x - x')^{n-m-s+q}$ which also vanishes for $x = x'$. Thus, the only term in the sum in (21) which is not zero is the one for which $m + s - q = n$. The integral I_2 thus is given by

$$I_2 = \sum_{s=0}^p \binom{p}{s} (-1)^{m+s} \binom{m+s}{m+s-n} (n!) \left[\frac{d^{p+m-n}}{dx^{p+m-n}} D_p(x) \right]_{x=x'}, \tag{22}$$

yielding

$$\begin{aligned}
M_n(x', \tau) &\sim (n!) \tau D_n(x') + \\
&\quad \frac{\tau^2}{2} \sum_{p=1}^{\infty} \sum_{m=1}^{\infty} (-1)^{p+m} D_m(x') \sum_{s=0}^p \binom{p}{s} (-1)^{m+s} (n!) \binom{m+s}{m+s-n} \left(\frac{d}{dx'} \right)^{p+m-n} D_p(x') \\
&= (n!) \tau D_n(x') + \frac{\tau^2}{2} \sum_{m=1}^{\infty} D_m(x') \left[\sum_{p=1}^{\infty} \sum_{s=0}^p \frac{(-1)^{p+s} p!(m+s)!}{s!(p-s)!(m+s-n)!} \left(\frac{d}{dx'} \right)^{p+m-n} D_p(x') \right].
\end{aligned}$$

The last derivative within parenthesis is of a non-negative order, i.e., $p \geq n - m$. Moreover, factorials are of non-negative integers, i.e., $s \geq n - m$. With both these conditions one arrives at a final approximation of each conditional moment as a function of the Kramers–Moyal coefficients and their derivatives, namely

$$M_n(x', \tau) \sim (n!) \tau D_n(x') + \frac{\tau^2}{2} \sum_{m=1}^{\infty} D_m(x') \sum_{p=p_0}^{\infty} \sum_{s=s_0}^p \frac{(-1)^{p+s} p!(m+s)!}{s!(p-s)!(m+s-n)!} \left(\frac{d}{dx'} \right)^{p+m-n} D_p(x'), \quad (23)$$

with $p_0 = \max(1, n - m)$ and $s_0 = \max(0, n - m)$. Eq. (23) yields the equations (13a) and (13b) in Ref. (Gottschall and Peinke 2008) for $n = 1$ and $n = 2$ respectively.

Eq. (23) holds a set of equations relating the conditional moments as functions of the Kramers–Moyal coefficients and their respective derivatives. In practice, one computes numerically the conditional moments and from it estimates the Kramers–Moyal coefficients. However, inverting Eq. (23) is not feasible, and a further approximation is required. Therefore, similarly to what was done for the second-order correction of the first two Kramers–Moyal coefficients (Gottschall and Peinke 2008; Rinn *et al.* 2016), we approximate Eq. (23) neglecting terms having derivatives, which implies $p + m - n = 0$, i.e., $p = n - m$. Furthermore, since $p \geq \max(1, n - m)$ and $p \geq s \geq \max(0, n - m)$, one has additionally $m \leq n - 1$ and $s = n - m$ respectively. Introducing these conditions in Eq. (23) yields our final approximation:

$$M_n(x', \tau) \sim (n!) \tau D_n(x') + \frac{(n!) \tau^2}{2} \sum_{m=1}^{n-1} D_m(x') D_{n-m}(x'). \quad (24)$$

Notice that the approximation of the conditional moments, as given in Eq. (24), has the practical advantage of expressing the conditional moment of order n -th from the Kramers–Moyal coefficients up to order n , which enables to compute recursively the Kramers–Moyal coefficients from the numerical computation of the conditional moments.

For jump processes we will need to estimate the first six Kramers–Moyal coefficients (Anvari *et al.* 2016b), which, from Eq. (24), read

$$M_1(x, \tau) = \tau D_1(x), \quad (25a)$$

$$M_2(x, \tau) = 2\tau D_2(x) + \tau^2 D_1^2(x), \quad (25b)$$

$$M_3(x, \tau) = 6\tau D_3(x) + 6\tau^2 D_1(x) D_2(x), \quad (25c)$$

$$M_4(x, \tau) = 24\tau D_4(x) + 12\tau^2 (2D_1(x) D_3(x) + D_2^2(x)), \quad (25d)$$

$$M_5(x, \tau) = 120\tau D_5(x) + 120\tau^2 (D_1(x) D_4(x) + D_2(x) D_3(x)), \quad (25e)$$

$$M_6(x, \tau) = 720\tau D_6(x) + 360\tau^2 (2D_1(x) D_5(x) + 2D_2(x) D_4(x) + D_3^2(x)). \quad (25f)$$

Finally, inverting Eqs. (25) yields Eqs. (14).

The formulas for the relations Eq. (25) and the corrections F_n given by Eq. (14) are given in symbolic python by the function `M_formula` and `F_formula`, respectively. This numerical procedure is generalised to any maximal order N needed for the data analysis. For jump-diffusion processes $N = 6$ is sufficient.

B. Higher-order corrections and Kramers–Moyal coefficients

A more accurate approximation is possible by combining Eqs. (23) and (24). More precisely, the steps to solve numerically Eq. (23) are the following ones:

- Solve Eqs. (14) introducing the conditional moments, extracted directly from the data, up to order N , and derive the Kramers–Moyal coefficients $D_n(x)$ as in Eq. (13).
- Compute the derivatives up to order N_d (see discussion below).
- Introduce the derivatives of the Kramers–Moyal coefficients and the empirical conditional moments in Eq. (23) and solve it with respect to the Kramers–Moyal coefficients.
- Repeat steps S1 and S2 until the Kramers–Moyal coefficients converge within a pre-given numerical accuracy.

Moreover, since the Kramers–Moyal coefficients are typically polynomials of lower order, not larger than five or six, the derivative order N_d is a finite number, which leads to a simplification of Eq. (23). Namely, the derivative in the sum obeys $0 \leq p + m - n \leq N_d$. Thus, $p_0 \leq p \leq N_d + n - m$. Since $p_0 = \max(1, n - m)$ one has $N_d + n - m \geq 1$ and therefore the sum over m is bounded by $1 \leq m \leq N_d + n - 1$. Since $N_d > 1$, $p_0 = 1$ and therefore the sum over p is also bounded by $1 \leq p \leq N_d + n - m$.

Introducing these bounds in Eq. (23) and following steps S0-S3 above, yields the second-order approximation of the Kramers–Moyal coefficients.

Lastly, we present a general framework for obtaining all moments and Kramers–Moyal coefficients. Eq. (24) can be written as

$$\begin{aligned} M_n(x', \tau) &\sim (n!) \tau \hat{B}_{n,1}(D_1(x'), D_2(x'), \dots, D_n(x')) \\ &+ \frac{(n!) \tau^2}{2} \hat{B}_{n,2}(D_1(x'), D_2(x'), \dots, D_{n-1}(x')) , \end{aligned} \quad (26)$$

where $\hat{B}_{n,2}$ are ordinary Bell polynomials, given by

$$\hat{B}_{n,k}(x_1, x_2, \dots, x_{n-k+1}) = \sum \frac{k!}{j_1! j_2! \dots j_{n-k+1}!} x_1^{j_1} x_2^{j_2} \dots x_{n-k+1}^{j_{n-k+1}} . \quad (27)$$

In the case of Eq. (26) we have $k = 2$.

The ordinary Bell's polynomials fulfil a reciprocal relation, namely any sum of ordinary Bell's polynomials of the form

$$y_n = \sum_{k=1}^n B_{n,k}(x_1, \dots, x_{n-k+1}) \quad (28)$$

can be inverted as

$$x_n = \sum_{k=1}^n (-1)^{k-1} (k-1)! B_{n,k}(y_1, \dots, y_{n-k+1}). \quad (29)$$

Consequently, substituting x_n and y_n by M_n and D_n , respectively, and applying Eq. (29), the inverse relation for D_n in Eq. (26) reads

$$D_n(x') = \frac{1}{\tau(n!)} \left[\hat{B}_{n,1}(M_1(x', \tau), M_2(x', \tau), \dots, M_n(x', \tau)) - \frac{\tau}{2} \hat{B}_{n,2}(M_1(x', \tau), M_2(x', \tau), \dots, M_{n-1}(x', \tau)) \right]. \quad (30)$$

This relation simply requires an iterative process for obtaining the $n-1$ conditional moments to retrieve the Kramers–Moyal coefficient D_n , which is computationally inexpensive.

Affiliation:

Leonardo Rydin Gorjão, Dirk Witthaut

Forschungszentrum Jülich, Institute for Energy and Climate Research – Systems Analysis and Technology Evaluation (IEK-STE), 52428 Jülich, Germany

and

Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany

Pedro G. Lind

Department of Computer Science, OsloMet – Oslo Metropolitan University, P.O. Box 4 St. Olavs plass, N-0130 Oslo, Norway

and

Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal

2.3.3 Publication #8

L. Rydin Gorjão, K. Riechers, F. Hassanibesheli, D. Witthaut, and P. G. Lind, under the working title *Dansgaard-Oeschger events: Change in stability and jumps modelled via univariate and bivariate jump-diffusion processes*, Ref. [8].

Status: in preparation

Dansgaard–Oeschger events: Change in stability and jumps modelled via univariate and bivariate jump-diffusion processes

Leonardo Rydin Gorjão,^{1,2} Keno Riechers,³ Forough Hassanibesheli,^{4,3} Dirk Witthaut,^{1,2} and Pedro G. Lind^{5,6}

¹*Forschungszentrum Jülich, Institute for Energy and Climate Research - Systems Analysis and Technology Evaluation (IEK-STE), 52428 Jülich, Germany*

²*Institute for Theoretical Physics, University of Cologne, 50937 Köln, Germany*

³*Research Domain IV – Complexity Science, Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany*

⁴*Department of Physics, Humboldt-Universität zu Berlin, Newtonstraße 15, 12489 Berlin, Germany*

⁵*Department of Computer Science, OsloMet – Oslo Metropolitan University, P.O. Box 4 St. Olavs plass, N-0130 Oslo, Norway*

⁶*Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Av. das Forças Armadas, 1649-026 Lisboa, Portugal*

Dansgaard–Oeschger (D–O) events are sudden climatic transitions observed during the last glacial period, which one finds across recordings of oxygen-18 ($\delta^{18}O$) and dust counts in Greenland’s ice sheet, in the GRIP and NGRIP recording projects. The transition of states is attributed to a bistability of potential of the recordings, which could lead to rapid changes as the climatic variables hop between the minima of the potential. In this article we employ a data-driven analysis of the recording under the purview of stochastic processes and show that: 1) there is a change from a bistable to a unstable potential of the dust count, via an imperfect supercritical pitchfork bifurcation. 2) the $\delta^{18}O$ recording is discontinuous and thus best modelled via a jump-diffusion model. We present a bivariate jump-diffusion model that further indicated there is no coupling between the diffusion and jump variables, leaving only room for a coupling between the drift functions of the $\delta^{18}O$ and dust variables.

I. INTRODUCTION

Dansgaard–Oeschger (D–O) events are abrupt transition of the northern hemisphere temperature, seen across paleo-climatic records from the past 100 000 years [1]. These abrupt transitions can result in changes of over 6° Celsius of the temperature in a span of less than 40 years and are visible across several proxy temperature records from the Last Glacial Period. These events are believed to be Poisson distributed [2, 3], superseding previous theory that purported a periodicity of roughly 1 470 years of the events [4]. Recent hypotheses supports that D–O events emerge from a coupling effect between ocean and atmosphere dynamics, possibly linked to the Atlantic meridional overturning circulation [5, 6].

A well established method to model and examine D–O events relies on stochastic modelling of paleo-climate processes, which have seen successful application [7], even including non-Markovian, i.e., memory in the modelling [8], or explicit delayed coupling [9]. Commonly used measurements are the concentration of the oxygen-18 isotope, denoted $\delta^{18}O$, and the concentration of dust in the ice-core recordings, which are proxies of the surface air temperature and large-scale atmospheric circulation changes, respectively. They describe local surface temperature changes, in the former, and global atmospheric circulation, in the latter. These seemingly distinct phenomena show a large degree of correlation [10, 11].

In this article, we seek to extend the simple stochastic models to include discontinuous trajectories, alongside the common diffusion behaviour seen in the data, to model $\delta^{18}O$ and the concentration of dust. D–O events

are generically described as very abrupt transitions—i.e., jumps—in the temperature of the northern hemisphere. Previous stochastic models describe these changes by modelling the aforementioned variable with regime-switching models or bistable potentials, which leads to fast but nevertheless continuous transitions between states. The aim of this work is three-fold. First, include explicitly discontinuous jumps to account for the fast D–O transitions. Then include and analyse a potential coupling of the $\delta^{18}O$ and dust recordings. Finally, extract non-parametrically the functional forms of the parameters underlying the stochastic process—i.e., the drift functional, the diffusion, and the jumps,

This article is structured as follows. In Sec. II we introduce the Kramers–Moyal expansion as the prime method to extract the parameters of a stochastic model from recorded time series. The higher-order terms in the expansion provide a distinguished test for the continuity of the underlying process. Our analysis suggests that the dynamics of $\delta^{18}O$, unlike the dynamics of the dust recordings, cannot be modelled as a continuous stochastic process, but should include jumps explicitly. In Sec. III, we establish a generalised stochastic approach for the analysis of Dansgaard–Oeschger (D–O) events under the purview of jump-diffusion models. We present first a univariate analysis of the recordings and extract the jump rate and jump amplitude of the $\delta^{18}O$ recordings. We then extend the model to a bivariate (two-dimensional) jump-diffusion model to account for a potential coupling of the stochastic variables $\delta^{18}O$ and dust count. We extract non-parametrically the parameters of the process, showing that there is no coupling in the diffusion or jump

components of the model. Analysing of the drift functions, we find that the dust count recordings exhibits a change in stability dependent on the $\delta^{18}O$ recordings, changing from bistability to unistability, via an imperfect supercritical pitchfork bifurcation.

II. KRAMERS–MOYAL EXPANSION OF DISCONTINUOUS STOCHASTIC PROCESSES

In this section we will motivate an analysis of the dynamics of $\delta^{18}O$ and the dust count recordings under the purview of stochastic processes. We introduce the Kramers–Moyal expansion as the prime method to extract stochastic parameters from the data. We show that the $\delta^{18}O$ recordings exhibits elevated values for higher Kramers–Moyal coefficient, suggesting the presence of jumps or discontinuities in the underlying process. These finding form the basis for the development of a stochastic model in the next section.

A. Data and data pre-processing

Our analysis is based on the recordings of oxygen-18 ($\delta^{18}O$) and dust counts in Greenland’s ice sheet provided by the GRIP and NGRIP projects, as displayed in Fig. 1. Boers *et al.* [9] motivate a pre-processing of the signals, which we implement here with a caveat. Firstly, recall that the ice measurements are taken at a fixed 5 cm cones, which, although spatially uniform, are not temporally uniform. We interpolate the data to an equidistant time axis of 5 year intervals. Secondly, we fill the missing data with a next-neighbour interpolation. Finally, we apply a Butterworth low pass filter of fourth order with a cutoff frequency of 0.02, i.e., 50 years. The final step comes with one caveat: the low-pass filter creates artificial correlations in the increments of the data (see App. A), which affects the stochastic analysis. For this reason we will work with both the “raw” and the “processed” data in the following. Later in Sec. II, and throughout Sec. III to the end of the paper we will solely use the “raw” data, and the presence of discontinuities becomes central to the model presented below.

B. Formal definitions

A stochastic process is the mathematical description of the dynamics of a variable $X(t)$ subject to random influences. More precisely, a stochastic process is a mapping from time $t \in \mathbb{R}$ to the random variables $X(t)$ in some adequate state space. Stochastic processes can be either analysed in terms of the random variables following a stochastic differential equations or in terms of the their probability density function $p(x, t)$ following a partial differential equation. The connection between these

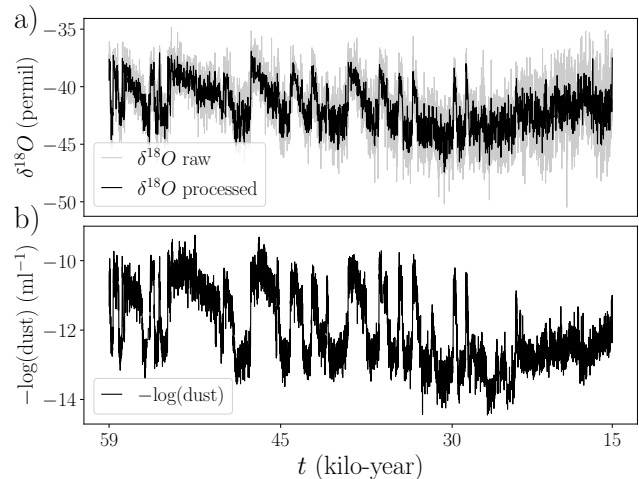


FIG. 1. a) Recordings of the $\delta^{18}O$ measurements in Greenland’s icesheet, raw and processed. b) Recordings (in logarithmic scale) of dust in Greenland’s icesheet. Data can be found in Ref. [12, 13]

two descriptions is given by the Kramers–Moyal expansion.

The Kramers–Moyal equation, stemming from the eponymous expansion, of the conditional probability density function $p(x, t+\tau|x', t)$ is the partial differential equation given by

$$\frac{\partial}{\partial t}p(x, t+\tau|x', t) = \sum_{m=1}^{\infty} \left(-\frac{\partial}{\partial x}\right)^m D_m(x)p(x, t+\tau|x', t). \quad (1)$$

If the process is sufficiently continuous, the third and higher order terms vanish, which directly leads directly to the Fokker–Planck equation (forward Kolmogorov or Smoluchowski equation) for the conditional probability $p(x, t+\tau|x', t)$ given by

$$\begin{aligned} \frac{\partial}{\partial t}p(x, t+\tau|x', t) &= \frac{\partial}{\partial x}D_1(x)p(x, t+\tau|x', t) \\ &+ \frac{\partial^2}{\partial x^2}D_2(x)p(x, t+\tau|x', t). \end{aligned} \quad (2)$$

We will later show that this is insufficient to represent the evolution process of $\delta^{18}O$ recordings.

Now to retrieve the Kramers–Moyal coefficients strictly from data we evaluate the transition probability densities in the limit of a vanishing time step $\tau \rightarrow 0$ (numerically we consider the shortest increment in the data),

$$\begin{aligned} D_m(x) &= \frac{1}{m!} \lim_{\tau \rightarrow 0} \frac{M_m(x, \tau)}{\tau} \\ &= \frac{1}{m!} \lim_{\tau \rightarrow 0} \frac{1}{\tau} \langle (x(t+\tau) - x(t))^m |_{x(t)=x} \rangle, \end{aligned}$$

in which above we employ a Nadaraya–Watson estimator with an Epanechnikov kernel [14–16].

Under the purview of the Kramers–Moyal expansion we analyse both the $\delta^{18}O$ and the dust count (in a logarithmic scale, as seen in Fig. 1) to uncover a formal

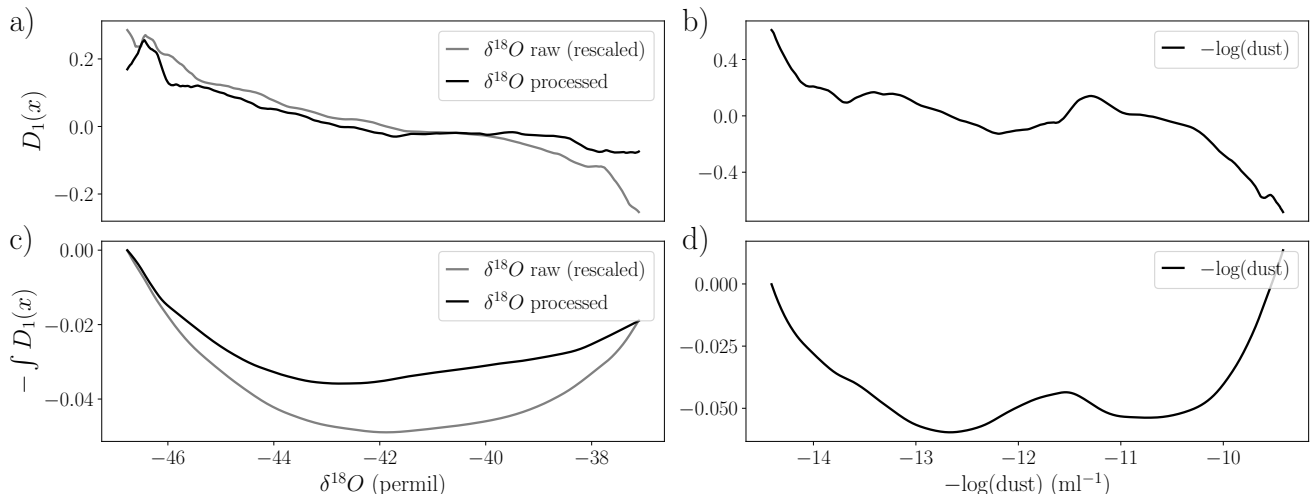


FIG. 2. a) Drift coefficient $D_1(x)$ of both the processes $\delta^{18}O$ recordings and the raw (unprocessed) $\delta^{18}O$ recordings (divided by 20 to match the scale). c) The potential well associated with $\delta^{18}O$ recordings, which shows a single minimum. Similarly, b) displays the drift coefficient $D_1(x)$ of the dust count, and d) the associate potential well. Notice that this suggest the dust count could be modelled via a double-well potential, but less so the dynamics of the $\delta^{18}O$.

structure of the parameters of the underlying stochastic processes. Let us now evaluate the first, second, and fourth Kramers–Moyal coefficient in a univariate (one-dimensional) setting and discuss the existence of mono and bistable potentials before moving to a bivariate setting in Sec. III.

C. Univariate stability of the $\delta^{18}O$ and dust count

The first Kramers–Moyal coefficients $D_1(x)$ relates to a stochastic process drift, i.e., the nature of the process to follow a deterministic behaviour, sometimes called mean-reverting strength. In the physics literature this relates directly to the potential energy, if one is to think of the described process as a particle moving in a given potential well.

In Fig. 2 we display the drift coefficient and the associated potential well, i.e., the integral over the drift $-\int D_1(x)dx$ (integration constant apart). In case of the dust count, the integral appears as a double-well potential, suggesting a bistable dynamics. This finding is not surprising in view of the rapid changes between different regimes observed in the time series in Fig. 1. Remarkably, no such bistability is found for the $\delta^{18}O$ data despite the existence of rapid changes in the trajectory. If one considers either the raw or the processed data, both seem to account for a single well (linear drift) of the $\delta^{18}O$ recording.

Here we emphasise two important issues: first, the presence of a bimodal distribution of the probability density function $p(x, t)$ of the recordings is not a sufficient argument that the drift (i.e., potential) is bistable. Bistability can be achieved variously via complicated diffusion functions, or generally with more complicated form of noise

(e.g. Lévy-like noise) [17–19]. Secondly, we note here that in the subsequent analysis in Sec. III we will show that the bistability of the dust count *is explicitly dependent* on the $\delta^{18}O$ recordings, which is impossible to judge from the univariate analysis presented thus far.

Relevantly, this does not preclude the $\delta^{18}O$ recordings from seemingly having two states of existence, and henceforward we will argue that the dynamical driver for the apparent two states—obvious and justified in the dust count—is riddled with jumps, possibly induced from the changes of state in the dust count.

Before we do so, let us evaluate, for the sake on completion, the second Kramers–Moyal coefficient $D_2(x)$ to show that a constant value is sufficient to describe the noise amplitude, which excludes the necessity of state-dependent noise. More importantly, we need to evaluate the second Kramers–Moyal coefficient $D_2(x)$ in contrast with the fourth Kramers–Moyal coefficient $D_4(x)$ to motivate the presence of jumps in the data, which we do below.

D. Studying the diffusion of the processes

In Fig. 3 we display the second Kramers–Moyal coefficient $D_2(x)$ as a function of the respective state variable x . We have included direct results via Eq. (3) alongside with a set of corrections proposed in Ref. [20] and described in detail in the App. B. Our main finding is that the fluctuation (noise) strength is constant to good approximation, i.e., it does not depend on the value of $\delta^{18}O$ or dust count, respectively.

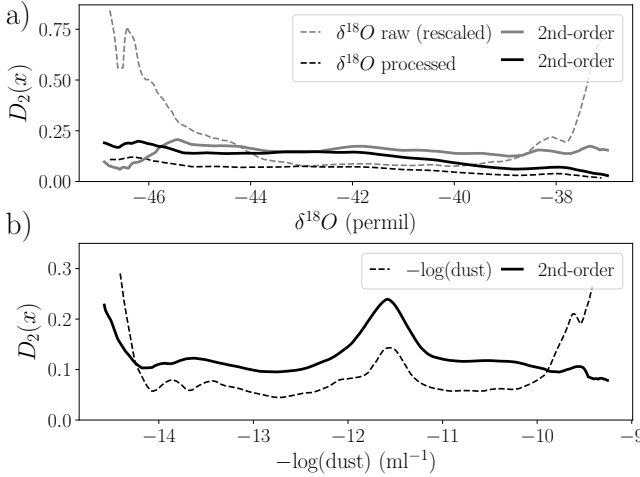


FIG. 3. Diffusion coefficients $D_2(x)$ of a) $\delta^{18}\text{O}$ measurements and b) dust (in logarithmic scale) in Greenland’s icesheet. Include as well are a set of corrections relevant for low-sampled recording [20], discussed in App. B. For both recordings (solid lines) the respective diffusion coefficients can be taken as constant. The protrusion in the centre of the $-\log(\text{dust})$ count in panel b) is the local maximum of the double-well potential (see Fig. 2 d)), where the number of recordings is minute. (the raw data is divide by 20 to match the scale).

E. The emergence of discontinuities in the $\delta^{18}\text{O}$ recording

Lastly, we examine the higher-order coefficient of the Kramers–Moyal coefficients, in particularly the fourth Kramers–Moyal coefficient $D_4(x)$. Higher-order coefficient of the Kramers–Moyal coefficients are vanishing for a regular continuous stochastic process, commonly referred to as Pawula’s theorem [21, 22]. In the presence of some discontinuous or jumpy processes these coefficients deviate from zero (in comparison with the first and second Kramers–Moyal coefficient). In this case the the Fokker–Planck equation (2) is no longer suited to describe the process and one needs to consider the full expansion given by Eq. (1). We note here that the presence of correlated forms of noise is also sufficient to generate higher-order Kramers–Moyal coefficients. However, we exclude this option as the auto-correlation of the increments of the data shows no correlations (apart from the shortest increment), see App. A). Internal correlations of the increments are possible (and likely), but that the sampling rate of our recordings is not sufficient to capture this, i.e., we are above the Einstein–Markov length [15], and a stochastic processes without correlations is an adequate description.

In Fig. 4 we display the fourth Kramers–Moyal coefficient $D_4(x)$ for the two processes in comparison to the second coefficient. The ratio $D_4(x)/D_2(x)$ shown in the insets provides a versatile indicator for the presence of vanishing of this higher order moment. Interestingly, the raw $\delta^{18}\text{O}$ recordings display a $D_4(x)$ of equivalent

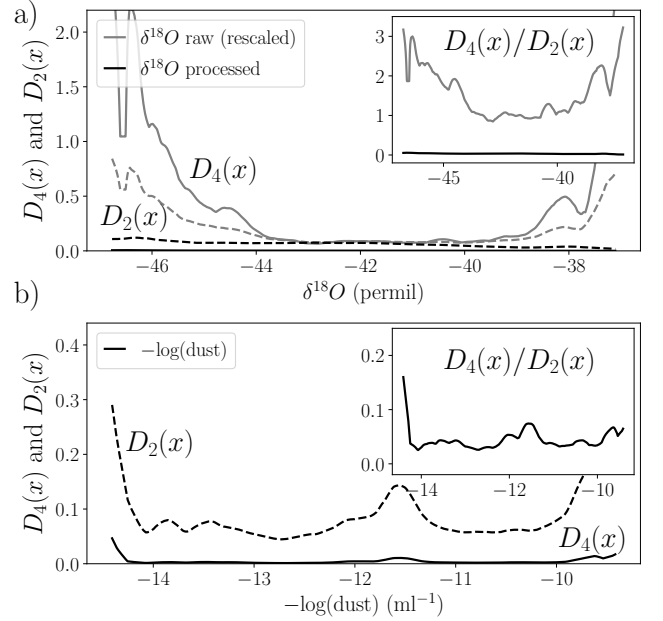


FIG. 4. Fourth-order Kramers–Moyal coefficient $D_4(x)$ of a) $\delta^{18}\text{O}$ measurements (raw and processed) and b) dust (in logarithmic scale) from the NGRIP Greenland’s icesheet recordings. The inset’s in each subplot compare $D_4(x)/D_2(x)$, a common ratio to evaluate if higher-order moments are vanishing. The raw $\delta^{18}\text{O}$ measurements display non-vanishing higher-order moments, suggesting the presence of discontinuities in the recordings.

magnitude to $D_2(x)$, suggesting that this process displays discontinuities. This is unsurprisingly not the case for the processed data since we applied a low-pass filter which quenches high frequencies, i.e., possible jumps in the data.

D–O event, seen as sudden transitions in the earth’s temperature, could possibly be related with this jumps in the concentration of $\delta^{18}\text{O}$. We will precise the rationale for the existence of discontinuous transitions in the following section.

F. Jumps in stochastic data

In Ref. [23], Lehnertz *et al.* study the comparative convergence of the moments of a stochastic process by evaluating their scaling with τ and propose a criterion to distinguish pure diffusion processes from processes with discontinuity, e.g., jump-diffusion processes. This criterion, denoted Q -ratio, is given by

$$Q(x, \tau) = \frac{M_6(x, \tau)}{5M_4(x, \tau)} \sim \begin{cases} \tau, & \text{for diffusions,} \\ c, & \text{for jumpy processes,} \end{cases} \quad (3)$$

where the moments $M_m(x, \tau)$ are given by Eq. (3). If the process is purely diffusive $Q(x, \tau) \sim \tau$ (i.e., a linear

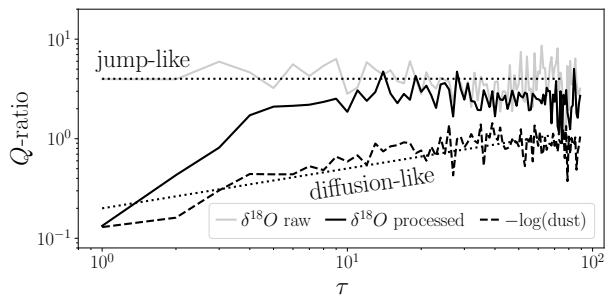


FIG. 5. Q -ratio of the $\delta^{18}\text{O}$ and the dust count, according to Eq. (3), in a double logarithmic scale. For the $\delta^{18}\text{O}$ one can observe the constant relation of $Q(x, \tau)$ with τ , indicating that, especially the raw recordings, are jumpy processes. The log dust count exhibits a linear relation with τ . The state x in $Q(x, \tau)$ is chosen at the maximum of the distribution of the recordings.

function of τ) and if the process exhibits discontinuous trajectories, $Q(x, \tau) \sim c$ (i.e., a constant over τ).

In Fig. 5 one can clearly see a constant relation of $Q(x, \tau)$ with τ for the raw $\delta^{18}\text{O}$ recordings, suggesting that this is a jump-like stochastic process. On the other hand, we observe a linear relation of $Q(x, \tau)$ with τ for the dust count, suggesting a purely diffusive process.

Having argued for the presence of discontinuities of the raw $\delta^{18}\text{O}$ recordings, we propose a two-fold augmentation of our analysis: First, we will introduce univariate jump-diffusion processes, which account for both a diffusion-like process, e.g. of a similar nature to a Langevin process, as well a jump-like components, which in our case is given by a Poisson distribution of the jumps. Indeed, the Poissonian distribution of the transitions seen in the $\delta^{18}\text{O}$ recordings have already been confirmed in Ref. [3]. Second, we will redesign our analysis into a bivariate case, i.e., we will study the two recordings in a two-dimensional setting, and introduced equivalently a bivariate jump-diffusion stochastic process [15, 16, 24].

III. UNIVARIATE AND BIVARIATE JUMP-DIFFUSION PROCESSES

Our previous results suggest including jumps explicitly in the analysis of D-O events. A large amount of models exist in the mathematical literature that include both a noise term $\sim dB$ and a jump-like element $\sim dJ$. Generally the problem surrounding these models is the ability to retrieve parameters strictly and non-parametrically from data, i.e., to be able to derive a model directly from the data. Jump-diffusion processes offer a direct relation between the partial differential representation of the evolution of the probability density function, through the Kramers–Moyal expansion in Eq. (1), and the parameters of the stochastic differential equation. Moreover, jump-diffusion processes produce exactly the commonly observed non-vanishing higher-order

Kramers–Moyal coefficients—just as seen in the recordings of $\delta^{18}\text{O}$ in Fig. 4.

Take a time-continuous Markov process, $X(t) \in \mathbb{R}$ given by three terms: one drift term, one diffusive term, and one Poissonian (jump) term. The evolution equation is given by

$$dX(t) = a(x, t) dt + b(x, t) dB(t) + \xi dJ(t), \quad (4)$$

where $a(x, t)$ is the drift strength, $b(x, t)$ is the diffusion strength or volatility, $B(t)$ is a Wiener process (Brownian motion), and $J(t)$ is a time-homogeneous Poisson jump process with jump rate $\lambda(x, t)$ and jump amplitude ξ , which is normally distributed $\xi \sim \mathcal{N}(0, s)$, with a variance $s = \langle \xi^2 \rangle$ [15, 24, 25]. For the case we consider here, we restrict ourselves to the set of stationary processes, thus $a(x, t) = a(x)$, $b(x, t) = b(x)$, $s(x, t) = s$, and $\lambda(x, t) = \lambda$ do not depend on time. The jump terms considered are also not state dependent.

Using the Kramers–Moyal expansion, one can recover the parameters of the jump-diffusion process (4), as given in Refs. [15, 24]

$$\begin{aligned} D_1(x) &= a(x), \\ D_2(x) &= \frac{1}{2} [b^2(x) + \lambda s], \\ D_{2n}(x) &= \frac{\lambda s^n}{2^n (n!)}, \text{ for } n > 2. \end{aligned} \quad (5)$$

One can see here directly the impact of the jumps on the higher-order Kramers–Moyal coefficients, as well as the possibility to utilise these to extract the jump rate λ and jump amplitude s , for which we can employ the relations [24]:

$$s = \frac{6D_6(x)}{D_4(x)}, \quad \lambda = \frac{8D_4(x)}{s^2} = \frac{2D_4(x)^3}{9D_6(x)^2}. \quad (6)$$

In Fig. 6 the jump amplitude $s_{\delta^{18}\text{O}}$ of the raw $\delta^{18}\text{O}$ recordings is retrieved via Eq. (6), which results in $s_{\delta^{18}\text{O}} = 5.10 \pm 1.65$. A similar examination for the processed data results in a jump amplitude of $\hat{s}_{\delta^{18}\text{O}} = 0.13 \pm 0.02$. This is not surprising since the high frequencies of this recording have been quenched by the low-pass Butterworth filter employed, as already mentioned above. Hence, this result needs to be interpreted with care.

The ratio of the Kramers–Moyal coefficients D_4 to D_2 shown in Fig. 4 suggests the processed data has vanishing higher-order Kramers–Moyal ($m > 2$) coefficients, which renders invalid the extraction of a jump term for the processed data.

Using Eq. (6) we can also recover the jump rate $\lambda_{\delta^{18}\text{O}}$ of the raw recordings, yielding $\lambda_{\delta^{18}\text{O}} = 0.70 \pm 0.49$. This non-parametric extraction needs to be evaluated with care, as it is given by the ratio of cubic over quadratic Kramers–Moyal coefficients, and thus has a high uncertainty. This nevertheless suggests that a jump of height s ($2s$) has a probability of taking place approximately

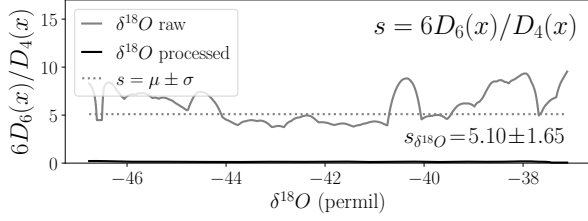


FIG. 6. Jump amplitude s is obtained by employing (6) for the raw $\delta^{18}\text{O}$ recordings, i.e., the ratio $6D_6/d_4$, from which one extract $s_{\delta^{18}\text{O}} = 5.10 \pm 1.65$ for the raw data. For comparison, the processed data, in which the high frequencies are quenched, results in $\hat{s}_{\delta^{18}\text{O}} = 0.13 \pm 0.02$ (this last result should be interpreted with care).

0.34 (0.12) times each 5 years, with s the variance of a Gaussian distribution with mean 0.

Having now extracted the jump rate and amplitude of the raw $\delta^{18}\text{O}$ recording we will motivate extending this analysis to a bivariate (two-dimensional) stochastic process, and in particular to a bivariate jump-diffusion process.

A. Why consider a bivariate (two-dimensional) jump-diffusion model?

A coupling of the $\delta^{18}\text{O}$ and the dust count has been posited as a possible explanation for the abrupt transitions in the recordings. The common method to model this is to coupled the variables directly in the drift, and possibly in the diffusion terms. This although precludes a coupling of the noises of the processes. In the following we will employ a bivariate (two-dimensional) analysis of the timeseries, assuming the existence of jumps in the system, thus offering *ab initio* a model capable of including couplings between the variables in their different terms (drift, diffusion, and jumps), but without positing the actual coupling structures. We find that there is no coupling between the diffusion (noise) terms nor the jump terms of the recordings, leaving only a possible coupling in the drift functions. In fact, we find that there is a coupling of the drift function of the dust count to the $\delta^{18}\text{O}$, but not the opposite. Here we most emphasise that the analysis put forth below is not restricted to bivariate jump-diffusion processes. Analysing the two processes solely under the purview of a bivariate diffusion process (i.e., set $\xi = dJ = 0$) equivalently leads to the conclusion that there are no couplings between the diffusion terms, and that the dust count is coupled to the $\delta^{18}\text{O}$, but not the opposite. This although naturally fails to capture the nature of the jumps of the $\delta^{18}\text{O}$ recordings.

Let us first introduce the bivariate jump-diffusion model [15, 16, 23, 24], which is simply an extension of the univariate case introduced in Eq. (4), with couplings between the diffusion (noise) terms dB_1 and dB_2 and jump terms dJ_1 and dJ_2 . The bivariate jump-diffusion

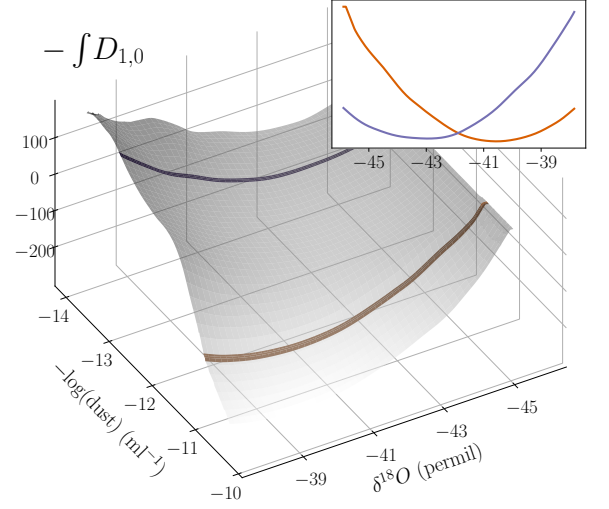


FIG. 7. Two-dimensional potential of the Kramers–Moyal coefficient $D_{1,0}$, which is obtained by performing a numerical integration of $a_1(\mathbf{x})$. The drift of the raw $\delta^{18}\text{O}$ is always linear, with negative slope, i.e., the potential is a convex quadratic functions. Inset shows the potential associated $\delta^{18}\text{O}$ for the two indicated minima of the dust count according to Fig. 2. For both cases the potential has a unique minimum, but the minimum shifts with the value of dust.

model of a two-dimensional vector $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ given by

$$\underbrace{\begin{pmatrix} dx_1(t) \\ dx_2(t) \end{pmatrix}}_{\text{drift}} = \underbrace{\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}}_{\text{drift}} dt + \underbrace{\begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}}_{\text{diffusion}} \underbrace{\begin{pmatrix} dB_1 \\ dB_2 \end{pmatrix}}_{\text{diffusion}} + \underbrace{\begin{pmatrix} \xi_{1,1} & \xi_{1,2} \\ \xi_{2,1} & \xi_{2,2} \end{pmatrix}}_{\text{Poissonian jumps}} \underbrace{\begin{pmatrix} dJ_1 \\ dJ_2 \end{pmatrix}}_{\text{Poissonian jumps}}, \quad (7)$$

where the set of drift and diffusion coefficients $\mathbf{a} = \mathbf{a}(\mathbf{x})$ and $\mathbf{b} = \mathbf{b}(\mathbf{x})$ may be state dependent.

Particularly, the terms $b_{1,2}$ and $b_{2,1}$ account for the coupling of one of the noise dimension to the other, e.g., how the stochastic fluctuation of the dust count affect the $\delta^{18}\text{O}$ recordings is manifested by $b_{1,2}$ (converse follows for $b_{2,1}$). In the same manner, the jumps in each dimension of the process can be coupled. For the case at hand it seems manifest that the dust count does not exhibit jumps, as its univariate analysis shows a vanishing ratio of $D_4(x)/D_2(x)$ (see Fig. 4), so we will now try to justify a data-driven analysis to obtain a stochastic jump-diffusion process adequate for a bivariate process of the $\delta^{18}\text{O}$ and dust count.

In general, a Kramers–Moyal expansion is possible for this two-dimensional process (and for any general N -dimensional process), which requires redefining the

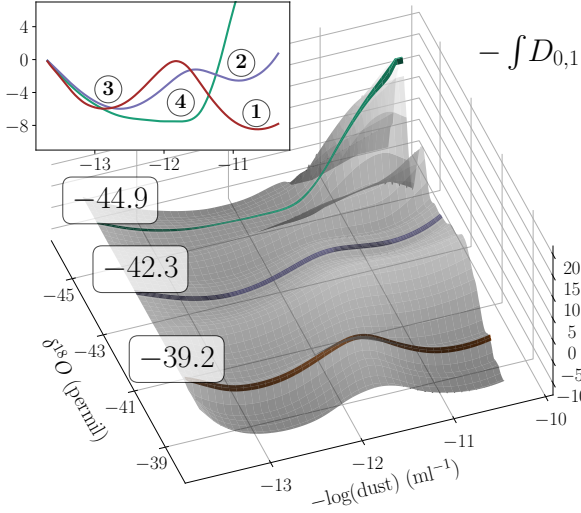


FIG. 8. The bivariate potential of the Kramers–Moyal coefficient $D_{0,1}$, which is obtained by performing a numerical integration of $a_2(\mathbf{x})$. Unlike the previous univariate case, which indicated that the dust count exhibited bistability, here we observe a set of different states, which is *explicitly dependent* on the value of $\delta^{18}O$. For low counts of $\delta^{18}O \sim -45$, the dust count seems to be uni-stable, but from approximately $\delta^{18}O \sim -42$ there is an emergence of a second minimum, which at approximately $\delta^{18}O \sim -39$ turns into the global minimum of the system. Numbers ① to ④ indicate the loss of bimodality of the $-\log(\text{dust})$ count with changing values of the raw $\delta^{18}O$. The inset show the three depicted potentials, exhibiting the change in stability.

Kramers–Moyal coefficients as two-dimensional scalar fields $D(\mathbf{x})_{a,b} = D(x_1, x_2)_{a,b}$, with $a, b \in \mathbb{N}$ the order of the coefficients. Notice here that for $a \neq b$ we generally have $D(\mathbf{x})_{a,b} \neq D(\mathbf{x})_{b,a}$.

Similarly to the one-dimensional process, there exists a set of relations between the Kramers–Moyal coefficients $D(\mathbf{x})$ and the parameters of Eq. (7), given by [24]

$$D_{1,0} = a_1, \quad D_{0,1} = a_2, \quad (8)$$

$$\begin{aligned} D_{1,1} &= b_{1,1}b_{2,1} + b_{1,2}b_{2,2}, \\ D_{2,0} &= \frac{1}{2} [b_{1,1}^2 + b_{1,2}^2 + s_{1,1}\lambda_1 + s_{1,2}\lambda_2], \\ D_{0,2} &= \frac{1}{2} [b_{2,1}^2 + b_{2,2}^2 + s_{2,1}\lambda_1 + s_{2,2}\lambda_2], \end{aligned} \quad (9)$$

$$\begin{aligned} D_{2,2} &= \frac{1}{4} [s_{1,1}s_{2,1}\lambda_1 + s_{1,2}s_{2,2}\lambda_2], \\ D_{4,0} &= \frac{1}{8} [s_{1,1}^2\lambda_1 + s_{1,2}^2\lambda_2], \\ D_{0,4} &= \frac{1}{8} [s_{2,1}^2\lambda_1 + s_{2,2}^2\lambda_2], \end{aligned} \quad (10)$$

where all higher order terms obey

$$D_{2\ell, 2m} = \frac{1}{2^\ell \ell!} \frac{1}{2^m m!} [s_{1,1}^\ell s_{2,1}^m \lambda_1 + s_{1,2}^\ell s_{2,2}^m \lambda_2]. \quad (11)$$

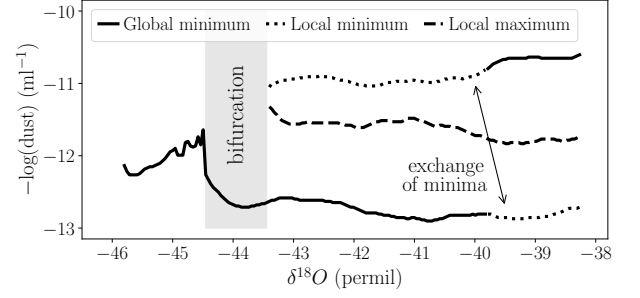


FIG. 9. Bifurcation of the potential of the dust count. By tracking the minima and maximum, i.e., the stable and unstable fixed points, respectively, we can study the bifurcation scheme of the potential of the dust count, which indicates that roughly around $\delta^{18}O \in [-44.5, -43.5]$ we observe an *imperfect supercritical pitchfork bifurcation*, giving rise to a pair of fixed points (one stable, one unstable). Transitions between these state can occur very rapidly, as the dust count can “hop” between the minima of the system, particularly close to the bifurcation point.

In the following sections we will justify, under this bivariate model, which parameters exist and which vanish, as well as which are state dependent or simply constant.

B. Simple linear coupling of the $\delta^{18}O$ with the dust count

We now evaluate the two-dimensional Kramers–Moyal coefficients numerically to elucidate the properties of the stochastic process. We start with the pivotal question of stability in the recording of $\delta^{18}O$.

Figure 7 shows the negative of the integral of the two-dimensional Kramers–Moyal coefficient $D_{1,0}$, which can be interpreted as a potential function for the dynamics of $\delta^{18}O$ as discussed above. The resulting potential is shown for all values of the dust count, highlighting the result for the two particular values $-\log(\text{dust}) = -11.2$ and $-\log(\text{dust}) = -13.3$. These two values have been obtained as the minima of the bistable dust potential in the univariate analysis in Fig. 2.

The reconstructed potential function shows a simple parabolic shape for all values of the dust count, corresponding to a linear drift term in the stochastic differential equation. However, the minimum of the parabola is strongly shifted as a function of the dust count. In conclusion, the data suggests a rather simple linear mean-reverting drift of $\delta^{18}O$, where the stable mean depends on the dust count.

C. Changing stability of the dust count

More interestingly, we can now re-examine the potential of the dust count, which is given as the negative of the integral over the Kramers–Moyal coefficient $D_{0,1}$. This

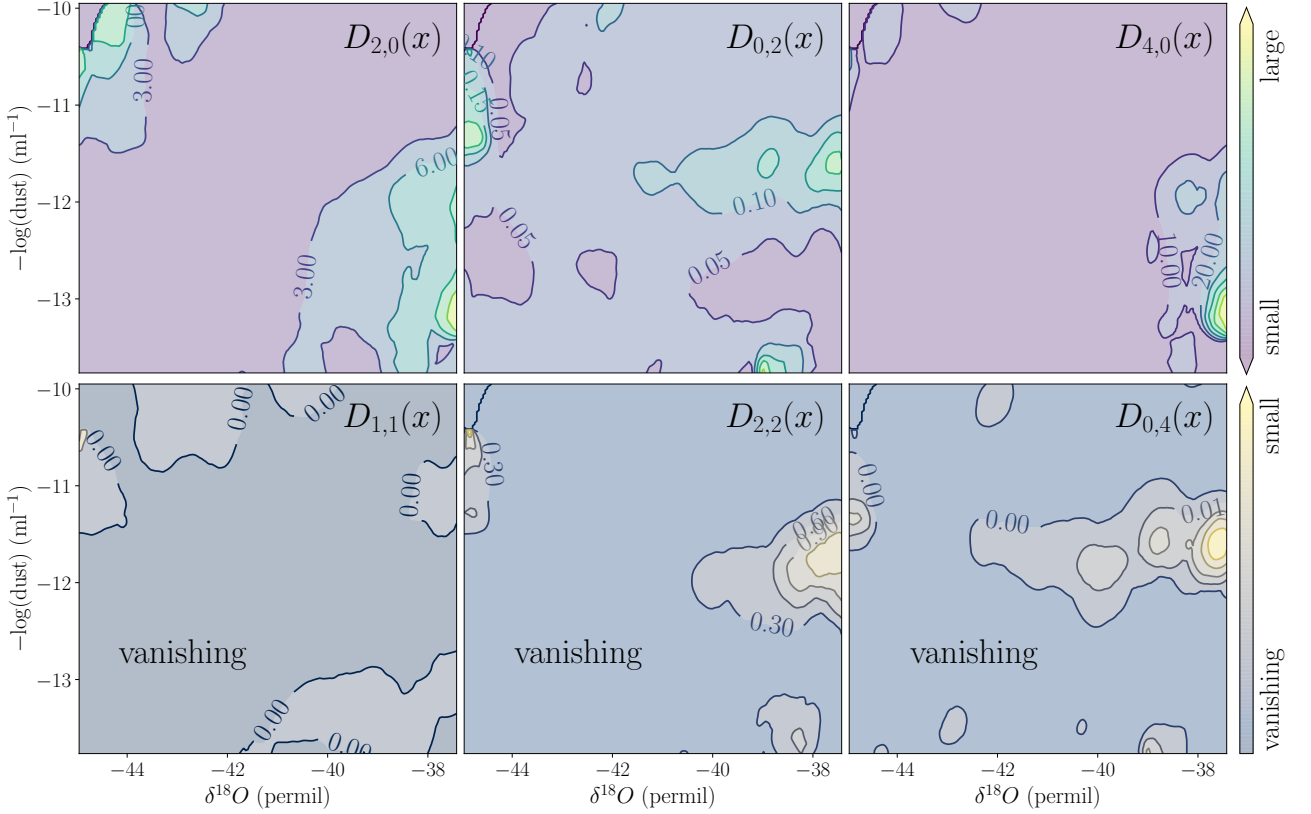


FIG. 10. Surfaces of the bivariate Kramers–Moyal coefficients $D_{1,1}$, $D_{2,2}$, and $D_{0,4}$, all of which vanish. This allows us to exclude, aided with our previous evaluation of vanishing jumps in the dust count, the terms $s_{2,1}$ and $s_{2,2}$ (or $\lambda_{2,\cdot}$). This one obtains evaluating $D_{2,2}$ and $D_{0,4}$, knowing ξ_1 is non-vanishing. Moreover, evaluating $D_{1,1} = 0$ equivalently implies either $b_{1,2}$ or $b_{2,1}$ are vanishing.

potential function reveal the existence of three different possible regimes for the dust dynamics, depending on the value of $\delta^{18}O$.

Depending on the value of $\delta^{18}O$, the potential of the $-\log(\text{dust})$ count changes from a bimodal to a unimodal regime, indicated by the numbers ① to ④. This change of the underlying potential shape justifies the changes in the trajectory of the $-\log(\text{dust})$ seen in Fig. 1. Moreover, the slow decay of the dust count after a transitions that increases its value to ~ -11 and then slowly returns to ~ -13 is explained by the flatness of the potential (in green), i.e., the change from state ③ to ④ takes place slowly. This subsequently leads the dust count to state ①, on the right side of the potential well.

From the potential seen in Fig. 8 we can extract the bifurcation diagram of the emergence of the fixed points, with $\delta^{18}O$ as the bifurcation parameter. In Fig. 9 we examine the local and global minima, i.e., the stable fixed points, potential in Fig. 8, as well as the local maximum, i.e., the unstable fixed point. The potential associated with the dust count undergoes an imperfect supercritical pitchfork bifurcation, somewhere in the range of $\delta^{18}O \in [-44.5, -43.5]$. This form of bifurcations are often associated with rapid changes of the variables, es-

pecially close to the bifurcation point, where the system can jump between the two minima. The potential around the minimum of the $-\log(\text{dust}) \sim -11$ becomes very shallow close to the bifurcation point, allowing the process to easily be “kicked out” of the local minimum when subject to fluctuations (induced by noise or inherent jumps).

D. Excluding noise and jump couplings in the timeseries

The bivariate jump-diffusion model presented in Eq. (7) has, in its full extent, twelve parameters. These include interaction between the diffusive noises dB_1 and dB_2 , mediated by the terms $b_{1,2}$ and $b_{2,1}$, as well as interactions between the jump terms dJ_1 and dJ_2 , mediated by $\xi_{1,2}$ and $\xi_{2,1}$. But in our example, while studying the univariate case of the dust count, we excluded the possibility of jumps (cf. Figs. 5 and 4). We will now in similar fashion refer to the relations in Eqs. (9), (10), and (11), and justify why several terms in our general formulation are non-existent.

In Fig. 10 we exhibit the Kramers–Moyal coefficients $D_{2,0}$, $D_{0,2}$, $D_{4,0}$, $D_{1,1}$, $D_{2,2}$, and $D_{0,4}$. The last three of

these vanish (bottom row of Fig. 10).

First, the vanishing values of $D_{0,4}$ and $D_{2,2}$ exclude the existence of jumps in the dust count recordings, as observed before for the univariate analysis. This sets $\xi_{2,2} = \lambda_2 = 0$. Moreover, this also set $s_{2,1} = 0$, under the relation that there are jumps in the $\delta^{18}O$ (i.e., $\lambda_1 > 0$).

Secondly, examining $D_{2,0}$ and $D_{0,2}$ indicates that neither of these two Kramers–Moyal coefficients vanish. For $D_{0,2}$ this implies that either or both $b_{1,1} > 0$ and $b_{2,1} > 0$. Under the purview that both these recordings are to some extent diffusion processes, they must have at least $b_{1,1} > 0$ and $b_{2,2} > 0$. Now if one combines this with $D_{1,1} = 0$ one has to obey that $b_{1,1}b_{2,1} + b_{1,2}b_{2,2} = 0$ (where $b_{i,j} > 0, \forall i, j$), thus enforcing that $b_{2,1} = b_{1,2} = 0$. Notice here that the result is identical even under $b_{1,1} = b_{2,2} = 0$, but this would just be a renaming of the $b_{i,j}$ terms ($b_{1,1} \rightarrow b_{1,2} = 0$ and $b_{2,2} \rightarrow b_{2,1} = 0$).

Thus we are left with a rather reduced set of relevant terms: Only $\delta^{18}O$ exhibits jumps ($s_{1,1} > 0$, $\lambda_1 > 0$, and $s_{2,1} = s_{2,2} = \lambda_2 = 0$, which also renders $s_{1,2}$ irrelevant). The diffusion terms $b_{i,j}$ do not show any coupling, as $D_{1,1} = 0$ ($b_{2,1} = b_{1,2} = 0$). The drift terms we have discussed before in Figs. 7 and 8, showing in the case of the dust count an explicit dependence on $\delta^{18}O$.

Here at last we need to emphasise another point. The couplings and correlations identified are not necessarily causation. The dependence of the dust count drift function on the $\delta^{18}O$ does not imply these two recordings are explicitly dependent as weather or climate variables, but instead that whatever coupling exists, these recordings serve as surrogates into the processes themselves coupled. This is as far as the argument can reach, but it does show quite clearly that the drift function of the $\delta^{18}O$ recordings remains pretty unaffected by the changes in the dust count, possibly indicating a one-directional coupling of these (or the underlying surrogate processes).

IV. CONCLUSION

In this article we have analysed the recordings of $\delta^{18}O$ and dust count from the NGRIP endeavour [1, 11]. We study these recordings under the purview of stochastic processes, in particular under jump-diffusion processes. We put forward an analysis that precludes any data processing of the recordings directly quenching slow or fast processes, i.e., low-pass filtering. Instead we study the “raw” data and find that the higher-order Kramers–Moyal coefficients are non-vanishing (contrary to what is expected under a Langevin or Fokker–Planck theory). This we posit is due to the presence of jumps (discontinuity) in the data, and further offer a justification by studying the scaling of the moments, known as Q -ratio, proposed by Lehnertz *et al.* in Ref. [23]. Here we note that the presence of jumps is only found in the $\delta^{18}O$ recordings, but not on the dust count. On the other hand, the dust count does seem convincingly well described by a double-well

potential.

These findings suggest an interpretation of the observed time series in terms of generalised stochastic models. We offer an explanation in terms of a bivariate jump-diffusion process with a simple constant coupling of the $\delta^{18}O$ to the dust count. On the other hand, the dust count does seem convincingly well described by a double-well potential.

In the bivariate analysis we observe a rich phenomenon of changing drift function of the dust count, supporting that this variable is explicitly dependent on the $\delta^{18}O$ (or whichever process this recording is surrogate for). We observe that the underlying potential undergoes an imperfect supercritical pitchfork bifurcation, i.e. a transition from a unistable to a bistable potential of the dust count.

Furthermore, from our bivariate analysis we exclude the existence of a coupling of the jump variables between both recordings ($s_{1,2} = s_{2,1} = 0$), supporting that the jumps observed in the $\delta^{18}O$ recordings do not seem to couple to the recordings of the dust count, neither that there are induced jumps in the dust count arising from the jumps present in the $\delta^{18}O$ recordings. The dust count also has no intrinsic jumps ($s_{2,2} = \lambda_2 = 0$), which matches what we observed previously from the univariate analysis of the data. We also show that there is no evidence of any noise coupling between the two recordings ($b_{1,2} = b_{2,1} = 0$), thus each variable has its independent noise function (B_1 and B_2). This leaves only room for a complex drift structure of the variables, most prominently of the dust count, as described above.

We hope this work puts forward a set of reasonable arguments for the existence of discontinuities (jumps) in the $\delta^{18}O$, which can find an interpretation under stochastic processes with jumps. Furthermore, bivariate analysis, as the one presented here, could serve as well to study other datasets, as they are a well-justified method to study coupling between variables, even if somewhat cumbersome to analyse. We hope that this mathematical analysis can also support climatologic studies, for instance on the existence and interpretation of early warning signs for D–O events [26] or the physics of multistability in the climate system [27].

ACKNOWLEDGMENTS

We gratefully acknowledge support from the German Federal Ministry of Education and Research (grant no. 03EK3055B) and the Helmholtz Association (via the joint initiative “Energy System 2050 – A Contribution of the Research Field Energy” and the grant “Uncertainty Quantification – From Data to Reliable Knowledge (UQ)” with grant no. ZT-I-0029). This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

-
- [1] S. O. Rasmussen, M. Bigler, S. P. Blockley, T. Blunier, S. L. Buchardt, H. B. Clausen, I. Cvijanovic, D. Dahl-Jensen, S. J. Johnsen, H. Fischer, V. Gkinis, M. Guille-
vic, W. Z. Hoek, J. J. Lowe, J. B. Pedro, T. Popp, I. K. Seierstad, J. P. Steffensen, A. M. Svensson, P. Val-
longa, B. M. Vinther, M. J. Walker, J. J. Wheatley, and M. Winstrup, A stratigraphic framework for abrupt cli-
matic changes during the Last Glacial period based on three synchronized greenland ice-core records: refining
and extending the INTIMATE event stratigraphy, *Qua-
ternary Science Reviews* **106**, 14 (2014).
- [2] P. D. Ditlevsen, M. S. Kristensen, and K. K. Andersen, The recurrence time of Dansgaard–Oeschger events and
limits on the possible periodic component, *Journal of Cli-
mate* **18**, 2594 (2005).
- [3] P. D. Ditlevsen, K. K. Andersen, and A. Svensson, The
DO-climate events are probably noise induced: statistical
investigation of the claimed 1470 years cycle, *Climate of
the Past* **3**, 129 (2007).
- [4] M. Schulz, On the 1470-year pacing of Dansgaard–
Oeschger warm events, *Paleoceanography* **17**, 4 (2002).
- [5] J. Lynch-Stieglitz, The atlantic meridional overturning
circulation and abrupt climate change, *Annual Review of
Marine Science* **9**, 83 (2017).
- [6] N. Boers, M. Ghil, and D.-D. Rousseau, Ocean cir-
culation, ice shelf, and sea ice interactions explain
dansgaard–oeschger cycles, *Proceedings of the National
Academy of Sciences* **115**, E11005 (2018).
- [7] D. Kondrashov, S. Kravtsov, A. W. Robertson, and
M. Ghil, A hierarchy of data-based ENSO models, *Jour-
nal of Climate* **18**, 4425 (2005).
- [8] D. Kondrashov, M. D. Chekroun, and M. Ghil, Data-
driven non-markovian closure models, *Physica D: Non-
linear Phenomena* **297**, 33 (2015).
- [9] N. Boers, M. D. Chekroun, H. Liu, D. Kondrashov, D.-D.
Rousseau, A. Svensson, M. Bigler, and M. Ghil, Inverse
stochastic–dynamic models for high-resolution Green-
land ice core records, *Earth System Dynamics* **8**, 1171
(2017).
- [10] S. J. Johnsen, H. B. Clausen, W. Dansgaard, N. S. Gun-
destrup, C. U. Hammer, U. Andersen, K. K. Ander-
sen, C. S. Hvidberg, D. Dahl-Jensen, J. P. Steffensen,
H. Shoji, Á. E. Sveinbjörnsdóttir, J. White, J. Jouzel,
and D. Fisher, The $\delta^{18}O$ record along the Greenland Ice
Core Project deep ice core and the problem of possible
Eemian climatic instability, *Journal of Geophysical Re-
search: Oceans* **102**, 26397 (1997).
- [11] U. Ruth, D. Wagenbach, J. P. Steffensen, and M. Bigler,
Continuous record of microparticle concentration and
size distribution in the central Greenland NGRIP ice core
during the last glacial period, *Journal of Geophysical Re-
search: Atmospheres* **108**, 10.1029/2002JD002376 (2003).
- [12] J. P. Steffensen, Centre for Ice and Climate, Niels Bohr
Institute, Data, icesamples and software (2014).
- [13] I. K. Seierstad, P. M. Abbott, M. Bigler, T. Blunier,
A. J. Bourne, E. Brook, S. L. Buchardt, C. Buizert,
H. B. Clausen, E. Cook, D. Dahl-Jensen, S. M. Davies,
M. Guillevic, S. J. Johnsen, D. S. Pedersen, T. J. Popp,
S. O. Rasmussen, J. P. Severinghaus, A. Svensson, and
B. M. Vinther, Consistently dated records from the
Greenland GRIP, GISP2 and NGRIP ice cores for the
past 104 ka reveal regional millennial-scale $\delta^{18}O$ gradi-
ents with possible Heinrich event imprint, *Quaternary
Science Reviews* **106**, 29 (2014).
- [14] D. Lamouroux and K. Lehnertz, Kernel-based regression
of drift and diffusion coefficients of stochastic processes,
Physics Letters A **373**, 3507 (2009).
- [15] M. R. R. Tabar, *Analysis and Data-Based Reconstruc-
tion of Complex Nonlinear Dynamical Systems* (Springer
International Publishing, 2019).
- [16] L. Rydin Gorjão, J. Heysel, K. Lehnertz, and M. R. R.
Tabar, Analysis and data-driven reconstruction of bi-
variate jump-diffusion processes, *Physical Review E* **100**,
062127 (2019).
- [17] A. V. Chechkin, J. Klafter, V. Y. Gonchar, R. Metzler,
and L. V. Tanatarov, Bifurcation, bimodality, and finite
variance in confined Lévy flights, *Physical Review E* **67**,
010102 (2003).
- [18] A. V. Chechkin, V. Y. Gonchar, J. Klafter, R. Metzler,
and L. V. Tanatarov, Lévy flights in a steep potential
well, *Journal of Statistical Physics* **115**, 1505 (2004).
- [19] R. Metzler and J. Klafter, The restaurant at the end of
the random walk: recent developments in the description
of anomalous transport by fractional dynamics, *Jour-
nal of Physics A: Mathematical and General* **37**, R161
(2004).
- [20] J. Gottschall and J. Peinke, On the definition and han-
dling of different drift and diffusion estimates, *New Jour-
nal of Physics* **10**, 083034 (2008).
- [21] R. F. Pawula, Generalizations and extensions of the
Fokker–Planck–Kolmogorov equations, *IEEE Transac-
tions on Information Theory* **13**, 33 (1967).
- [22] R. F. Pawula, Approximation of the linear Boltzmann
equation by the Fokker–Planck equation, *Physical Re-
view* **162**, 186 (1967).
- [23] K. Lehnertz, L. Zabawa, and M. R. R. Tabar, Charac-
terizing abrupt transitions in stochastic dynamics, *New
Journal of Physics* **20**, 113043 (2018).
- [24] M. Anvari, M. R. R. Tabar, J. Peinke, and K. Lehnertz,
Disentangling the stochastic behavior of complex time
series, *Scientific Reports* **6**, 35435 (2016).
- [25] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar,
Approaching complexity by stochastic methods: From bi-
ological systems to turbulence, *Physics Reports* **506**, 87
(2011).
- [26] N. Boers, Early-warning signals for Dansgaard–Oeschger
events in a high-resolution ice core record, *Nature com-
munications* **9**, 1 (2018).
- [27] M. Ghil and V. Lucarini, The physics of climate variabil-
ity and climate change, *Reviews of Modern Physics* **92**,
035002 (2020).
- [28] L. Rydin Gorjão, D. Witthaut, and P. G. Lind, JumpDiff:
A python library for statistical inference of jump-
diffusion processes in sets of measurements, *Forthcoming*
(2020).

Appendix A: Butterworth low-pass filter and correlations of increments

For the pre-processing of the data, discussed in Sec. II A, we introduced a method to reduce the noise of the $\delta^{18}O$ recordings by employing a Butterworth low-pass filter of fourth order, with cut-off period of 50 kiloyear. In Fig. 11 we plot the autocorrelation of the increments of the data, raw and processed, i.e., $\Delta(x(t)) = x(t+1) - x(t)$, at the shortest increment $t = 5$ years, to show that this form of filtering might be ill-advised, as it seems to induced in correlation of the data that might not exist.

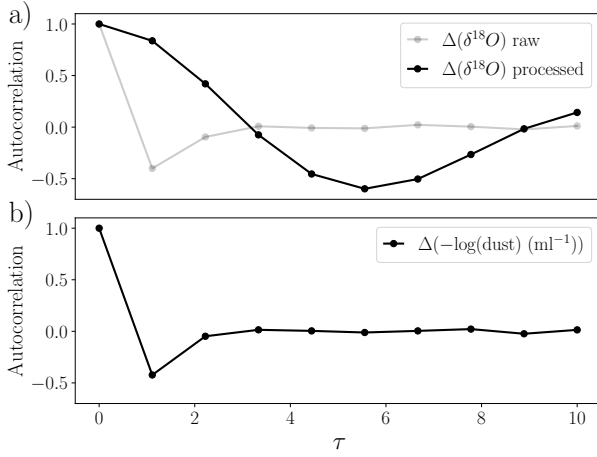


FIG. 11. Autocorrelation of the increments of both datasets. (a) displays the autocorrelation of the raw and processed $\delta^{18}O$ data, which indicated that the Butterworth low-pass filter introduces correlations on the increments. (b) autocorrelation of increments of the $-\log(\text{dust})$.

Appendix B: Second-order correction to the Kramers–Moyal operator

In order to correctly retrieve, from data, the Kramers–Moyal coefficients, we need to evaluate the operation in the Fokker–Planck equation Eq. (2). Let us focus on this equation for the moment, and rewrite it in a more formal

manner as an operator

$$\begin{aligned} \frac{\partial}{\partial t} p(x, t + \tau | x', t) &= \frac{\partial}{\partial x} D_1(x) p(x, t + \tau | x', t) \\ &\quad + \frac{\partial^2}{\partial x^2} D_2(x) p(x, t + \tau | x', t) \\ &= \mathcal{L}_{\text{FP}} p(x, t + \tau | x', t), \end{aligned} \quad (\text{B1})$$

with \mathcal{L}_{FP} the formal Fokker–Planck operator and

$$D_m(x) = \frac{1}{m!} \lim_{\tau \rightarrow \infty} \frac{M_m(x, \tau)}{\tau}, \quad (\text{B2})$$

where $M_m(x, \tau)$ is the m -order conditional moment, i.e.,

$$M_m(x, \tau) = \int_{-\infty}^{\infty} (x' - x)^m p(x', t + \tau | x, t) dx'. \quad (\text{B3})$$

which we introduced in Eq. (3) in a similar notation. In order to solve Eq. (2), one takes the formal step considering an initial conditions $\delta(x - x')$ as a starting point and employing the exponential representation of the operator, which we can decompose it into a power series as

$$\begin{aligned} p(x, t + \tau | x', t) &= \exp(\tau \mathcal{L}_{\text{FP}}) \delta(x - x') \\ &= \sum_{k=0}^{\infty} \frac{(\tau \mathcal{L}_{\text{KM}})^k}{k!} \delta(x - x'). \end{aligned} \quad (\text{B4})$$

From here we consider the 1st-order and 2nd-order approximation, i.e., truncation of the operation, as

$$\exp(\tau \mathcal{L}_{\text{FP}}) \sim 1 + \tau \mathcal{L}_{\text{FP}} + \frac{\tau^2}{2} \mathcal{L}_{\text{FP}} \mathcal{L}_{\text{FP}}. \quad (\text{B5})$$

Considering only the 1st-order, $\sim \tau$ we recovers the well-known relation between the conditional moments and the Kramers–Moyal coefficients, given by

$$D_m(x) = \lim_{\tau \rightarrow 0} \frac{M_m(x, \tau)}{(m!) \tau}. \quad (\text{B6})$$

If we now include the 2nd-order approximation, i.e., we consider terms up to $\sim \tau^2$, we obtain a corrective term for the second Kramers–Moyal coefficient

$$\begin{aligned} D_1(x) &= \lim_{\tau \rightarrow 0} \frac{1}{\tau} M_1(x, \tau), \\ D_2(x) &= \lim_{\tau \rightarrow 0} \frac{1}{2\tau} (M_2(x, \tau) - M_1(x, \tau)^2). \end{aligned}$$

We employ this correction to our examination solely to show that the diffusion coefficient, i.e., the amplitude of the fluctuations, is constant in space. In this work we extend our analysis beyond the Fokker–Planck equation to include higher-order terms, which elicits also the examination of considering second-order 2nd-order terms of Eq. (1), i.e., replacing \mathcal{L}_{FP} with \mathcal{L}_{KM} that represents correctly the operator in Eq. (1). This is addressed in a forthcoming publication [28].

Chapter 3

Conclusions

The energy transition and the mitigation of climate change are two of the central challenges of mankind in the 21st century. Mastering these challenges requires contributions and collaborations from all areas of science and technology, not least physics and scientific data analysis. The goal of this thesis is to contribute to the understanding of the dynamics of electric power systems and paleo-climate transitions. Hopefully, this work improves the understanding of these complex systems and that the developed software proves to be helpful for other researchers and engineers.

This chapter provides a short comprehensive discussion of the results of this thesis, grouped into three sections on power-grid frequency dynamics and stability, paleo-climatic events, and the software developed to analyse the respective data. The discussion will focus on overarching aspects of all studies and the connections to physics, whereas the domain specific conclusions and potential applications have already been discussed in some detail in the respective publications.

3.1 Power systems and power-grid frequency

In publications #1, #2, #4, and #5 a series of studies of power-grid frequency recordings are presented.

Results

The aforementioned studies include both a model-free statistical characterisation of power-grid frequency fluctuations as well as model-based studies elucidating the role of

different elements of the power system on the frequency. A comprehensive overview over the empirical distributions of grid around the world is provided, verifying a naturally occurring scaling phenomenon first proposed in Ref. [69]. The analysis of synchronous measurements at different locations provides insights into emergence of phase and amplitude synchronisation. A relaxation time constant has extracted, revealing a diffusive relation between the propagation of fluctuations and the distance, but a thorough theoretical understanding is yet to be developed. The existence of ubiquitous correlations in power-grid frequency recordings is revealed, suggesting a more sophisticated stochastic process with correlated elements is necessary in subsequent studies.

These works are relevant for the development of data-driven stochastic processes in the energy sciences. They provide an avenue to incorporate stochastic processes as the fundamental method of studying power-grid data, aided by the development of estimators for the drift and diffusion terms. As recordings of power-grid specific measures are usually scarce, dynamical models are often taken as the basis of an analysis in a bottom-up approach. These are integrated and tested against scarcely existent data, and naturally preclude the existence on intrinsic fluctuations, i.e., stochastic noise, in these systems.

The offered data science approach is based on an elementary Langevin-like approach to power-grid frequency dynamics. The characteristics of power-grid frequency data is examined with the aid of data-driven estimators, which extract solely from data the underlying fundamental characteristics of the data.

In publication #1 a Markovian Ornstein–Uhlenbeck process driven by a deterministic function representing the power imbalance is proposed as an elementary model for power-grid frequency fluctuations.

Strictly from a single recording of power-grid frequency data the fundamental parameters of the model are extracted: A short-term mean-reverting strength, equivalent to the primary control mechanism in power-grids is obtained via a non-parametric estimation of the drift coefficient; the long-term exponential relaxation of an Ornstein–Uhlenbeck is obtained by extracting the relaxation constant of such relaxation; the noise or intrinsic fluctuation terms are extracted via non-parametric estimation of the diffusion coefficient. The deterministic disturbances brought about by power imbalances are determined via the data science examination of their caused deviation from the nominal frequency.

In publication #2 a statistical study of power-grid frequency recordings is presented. The data shows that frequency fluctuations are not Gaussian distributed neither in itself nor in its increments, revealing rather large excess kurtosis in the latter. A superstatisti-

cal q -Gaussian is fitted to the data [75, 76], showing better results at capturing the heavy-tailedness of the data, yet no justification is offered for this particular model. A study of time-reversibility is also proposed by the analysis of two distinct three-point correlation functions, suggesting that the data is time-reversible. Furthermore, it is suggested that the process can be assumed as Markovian by utilising the Chapman–Kolmogorov test. This point in particular will be re-address in the discussion.

In publication #4 a first analysis of a large set of power-grid frequency recordings in different power grids around the world is complemented with a set of six synchronous recordings in the Central European grid. The natural scaling of fluctuation in these time-series is shown to scale as $1/\sqrt{N}$, with N the number of participants in a power-grid. This is well in line with scaling of any i.i.d. random variable with well defined mean and bounded variance, in accordance with the Lindeberg–Lévy central limit theorem. Subsequently the distribution of the increments of these timeseries is examined as a function of the time lag. Most scale-invariant stochastic processes exhibit non-Gaussian distribution of their increment statistics [77, 78], most often distributed as a superposition of log-normally distributed Gaussian processes [79, 80], or more generally some superstatistical formulation [75, 76]. Overall, all these distributions of incremental processes show heavy-tailed distributions, i.e., large kurtosis, which may or may not relax to Gaussian shaped distribution for increasing temporal lags. Both phenomena seem present in power-grid frequency recordings. Lastly, an examination of the fluctuations in six synchronous recordings in the same synchronous area is presented. A spatial analysis shows that fluctuations synchronise in a diffusive-like manner in space, i.e., they become identical between locations in a square relation with the distance between the locations.

In Publication #5 a thorough examination of a collection of six high-resolution synchronous power-grid recordings from the Nordic synchronous area is presented. In this work the aforementioned diffusive-like relation of amplitude synchronisation, first evidenced in Publication #4, is strengthened. This, moreover, is found to act as a stronger form of diffusion, i.e., as a superdiffusive processes as ones commonly seen in anomalous diffusion [81, 82], in a power relation with the Hurst coefficient of the underlying process. Additionally, the high temporal resolution of the data allows to analyse phase and amplitude synchronisation separately. This analysis suggest that phase synchronisation, i.e., the increase in correlation of the increments of the frequency recordings, occurs earlier than amplitude synchronisation and the time to reach synchrony between two locations scales only linearly with the distance.

Discussion

The aforementioned publications offer a novel perspective on power-grid dynamics, covering the main aspects of the topic: data analysis, physical interpretations, and modelling.

The obvious first point to discuss is the nature of the chosen model. It has become obvious throughout these investigations that the Markovian property is not strictly satisfied and any Markovian model can only serve as a first approximation to describe the dynamics of power-grid frequency. The application of the aforementioned Chapman–Kolmogorov test, stemming from the eponymous theorem, is not sufficient to discern whether one-dimensional stochastic processes are Markovian [83, 78]. The application of the Chapman–Kolmogorov test seems reliable solely at the level of increment statistics, not directly on data. This agrees well with a closer examination of a structure function, e.g., finite differences or log-returns, of the data, in the sense that it provides a necessary condition for Markovianity of the increments. It although fails as a verification that data—in particular, the increment statistics—is not Markovian. A closer examination of the increment statistics—in particular its covariance function—indicates without rebuttal that exponential-like correlations are present. For the suggested process, i.e., a simple Ornstein–Uhlenbeck process with non-multiplicative noise, one can be certain that the increment statistics reflects solely the nature of the noise.

The auto-correlation function is a natural tool to discern the Markovianity of a stochastic process, but much more precise insights can be obtained from the increment ratio statistics or power variations [84, 85]. Take a Ornstein–Uhlenbeck process with a drift function with bounded variation, fractional Gaussian noise with any Hurst (Hölder) index in $(0, 1)$, and a non-multiplicative diffusion (volatility) parameter [86, 87, 88]. A set of non-parametric estimators has been shown to exist that can precisely the Hurst (Hölder) index, the integral over the time-dependent diffusion strength (volatility), and the drift strength. A set of p^{th} -order k^{th} -difference operator, with $p > 1$, $k \geq 1$, which can be used as an estimator for the aforementioned stochastic parameters, was shown to converge in probability to these parameters, provided the process is stationary.

A direct application of these non-parametric estimator—much in contrast with the employed Nadaraya–Watson estimators—show directly that power-grid frequency does not obey the Markov property. This has important consequences for the modelling of stochastic processes but also for the methods used to analyse them. In a recent publication [89] it was shown that Nadaraya–Watson estimators are not adequate objects to study correlated processes as any choice of non-white noise renders these estimators

invalid. The proofs rely once more on the abnormal convergence in probability, which are shown to diverge from the actual parameters in non-Markovian stochastic differential equations. Nadaraya–Watson estimators rely solely on convergence in probability.

At this stage one must conclude that all Markovian approaches to the analysis and modelling of power-grid frequency dynamics must be regarded as a first approximation—elucidating some essential properties of the dynamics but failing to capture more intriguing aspects. This holds in particular for publication #2. Whereas all results on the distribution of frequencies and increments are still in good standing, the analysis of correlations is clearly oversimplified and should be corrected.

Note here that from a statistical point of view, at the distributional level most work is still in good standing. Take any correlated process with incremental which is Gaussian distributed, it does not invalidate nor alter in any fashion the observation of heavy platykurtic distributions at both the level of the power-grid frequency recordings or their increments (see, for example, the distributional properties of fractional Brownian motion). The examination offered in Publication #4 does not rely on the Markov property, except on the calculation of the scaling phenomena, which relies on the Nadaraya–Watson estimator. Nevertheless, a purely data science estimation is given in the Supplemental Material that does corroborates the findings.

Contemporaneous works

Power-grid frequency dynamics studies are not uncommon in the large scope of energy engineering [90], control theory [42, 91], nor by now the physics community [69, 92]. Yet, many studies are based on direct model analysis, in particular with emphasis on stability criteria [93, 42], graph-theoretical approaches [94, 95, 96], or numerical simulation [97, 98].

Data-driven approaches and real-world data analysis of power-grid frequency are scarce, with the notable exception of studies on intra and inter-area oscillations and their eigenfrequencies oscillations [99, 100, 101, 102] in the engineering community and a handful examples in the physics community [69, 92, 103, 104]. Real-world studies are difficult to perform given the lack of freely available data, which has seen critique from within the scientific community [105, 106].

One obvious obstacle to empirical studies of power-grid operation and stability is the lack of openly available data. Power grid operators closely monitor the grid and thus collect huge amounts of data, but are reluctant to share. This lack of data has been a

main motivation the effort to create an open data base [72, 4]. The importance of open data for energy science in general has been discussed and summarised in [105, 106].

As discussed repeatedly before, the power-grid frequency is intimately connected to the balance of generation and consumption of electric power. Frequency fluctuations are direct consequences of the stochastic fluctuations at the level of power generation and consumption. Recent studies have emphasised the importance of temporal correlations in wind turbine power generation from extensive real-world recordings [107]. The theoretical analysis of the recordings assumes fractional noise added on a cubic drift functional and incorporates a subsequent truncation for a lower and upper bound of power generation. This is perhaps the most daunting point as the choice of additive (fractional) noise is not bounded. The authors set a lower and upper cut-off power-generation, where possible more adequate stochastic models, as a Cox–Ingersoll–Ross model [108], general a Constant Elasticity of Variance (CEV) model [109], or more adequate Pearson diffusion processes [110] would render these effect without *ad hoc* criteria. The authors show that correlated noise can produce more adequate probability density functions than compared to plain uncorrelated Gaussian noises. They include as well an examination of the Hurst coefficient of the timeseries via detrended fluctuation analysis [111], thus justifying their choice of motion.

These results align particularly well with other examinations of temporal correlations of wind speeds [37], relegated to the supplemental information of the publication, where an examination of wind speed reveals likewise positively correlated phenomena. In Ref. [104] the authors study directly the detrended fluctuation analysis on scales larger than ten seconds, thus providing an analysis of the scaling phenomena at large scales, i.e., minutes, hours, days. The actual short-term memory properties of power-grid frequency are left unaddressed.

The particular structure of the underlying power-grid frequency fluctuations—in the purview of stochastic processes—has so far been conjectured to be a potential Lévy α -stable distribution, or, to account for different time scales in the process, a q -Gaussian distribution [69]. The first seems unlike, as it required unbounded, ill-defined variances, and thus must be seen with care. The second seems more in line with results in Ref. [104], including explicitly different time scales in the stochastic process. This, unfortunately, is not shown to a great extent, as the suggested q -Gaussian distribution are simply fitted to the probability density function of power-grid frequency recordings. An examination of the underlying structure function, i.e., the finite differences of power-grid frequency recordings, could unveiled the actual scale structure of the potential q -Gaussian-like

distributions [66]. Similarly, multifractal detrended fluctuation analysis [112] can be employed to unveil the scaling of the structure functions.

All of this combined forms a steady progress in adding to the comprehension of the stochastic nature of power-grid frequency dynamics, to which some of the publications present in this thesis contribute.

3.2 Bivariate jump-diffusion processes and Dansgaard–Oeschger events

Results

In publication #6 a standard model for univariate jump-diffusion processes proposed in Refs. [47, 67, 38] is extended to two dimensions

For the case of symmetric coupling in either the diffusive or jump terms, or for both, a close form linking the conditional moments and the parameters of the jump-diffusion process is possible. The actual solution of recovering the parameters is possible—within the limitations of their non-linearity and availability of data—and is it possible non-parametrically with the methods and software developed in publications #3 and #7.

In publication #8 bivariate jump-diffusion processes are utilised to model and analyse abrupt transitions in paleoclimatic records generally referred to as Dansgaard–Oeschger events. This framework provides a straightforward model for the abrupt transitions and allows to investigate the coupling or dependencies between two essential indicators of the climatic dynamics, the temperature and the concentration of dust in the atmosphere. The data analysis suggests the existence of an imperfect supercritical pitchfork bifurcation in the dust dynamics, which can be posited as one of the mechanisms behind the Dansgaard–Oeschger events.

Discussion

Application of data-driven stochastic models in paleo-climate data is not uncommon [62], yet most approaches fall under the category of pre-designed functional form for the drift and diffusion terms. Complex descriptions, involving generalised Langevin equations, offer a window into incorporating memory terms into these stochastic descriptions [113]. Yet, no approach with explicit discontinuous stochastic processes is known. Here note that bistable models always offer fast transitions between states, but these are not math-

ematically discontinuous transitions. A clear-cut distinction can be obtained by examining the structure function of the recordings, possible through, for example, spectral analysis, multifractal detrended fluctuation [112], or wavelet transform [114].

One noticeable difference in publication #8 to others in the field is the explicit absence of an *a priori* given model. In this work the use of non-parametric estimators eschew the need to propose a model prior to the data analysis—on the contrary, an examination of the estimators gives access to the functional form of a stochastic model’s parameters. Moreover, the particular augmentation of a general continuous-time diffusion (i.e., continuous stochastic process) to a discontinuous jump-diffusion process is not limiting. In opposition, by both examining the higher-order conditional moments of the timeseries as well as the method to distinguish jump-diffusions from pure diffusion from Ref. [68], the necessity to include a specific jump component becomes evident and well grounded.

A noticeable drawback—not intrinsic to the models but to the work in Publication #8—is the absence of an explicit memory kernel in the Langevin-like equation. Recent works have utilised generalised Langevin equations [113], but again always with *a priori* defined stochastic models, thus not providing a method to examine or extract the actual functional form of the memory kernel.

Contemporaneous works

Jump-diffusion processes are common in the mathematical literature in stochastic processes [115, 116]. Fast, discontinuous transitions are common in financial market data modelling [117, 118], commonly denoted market crashes. Poisson distributed jumps are perhaps the more common in the literature, but these *prima facie* naturally lead to an asymmetric, one-side distributions [119]. Poisson jump distributions with a Gaussian distributed amplitudes offer a neat and applicable model for symmetric distributed processes [120]. They naturally resemble symmetric Lévy α -stable distribution, yet have bounded variance [66]. A noteworthy pre-print publication addressing non-parametric estimators for Lévy driven processes could prove to be a breakthrough in the particularly difficult terrain of unbounded variance processes [65].

Jump-diffusion equations are much less common in the stochastic modelling of actual data sets than ordinary diffusion equations, while there is well established mathematical theory [120]. Two notable exceptions include the intermittency of solar power generation due to cloud coverage in Ref. [38], and the analysis of electroencephalogram timeseries

of brain activity [67, 68]. This comes hand-in-hand with the difficulty of disentangling the underlying parameters of timeseries from solely data-driven methods. This is aggravated with the interpretation of jumps in physical systems, which at first principle are assumed to be continuous. Interestingly, in Ref. [68] a first applicable criterion to distinguish purely diffusive and jump-diffusions is presented, denote Q -ratio. This, in a very similar fashion to the non-parametric Nadaraya–Watson estimators, relies on the scaling properties of the incremental timeseries for increasing finite-order differences. Fundamentally, a distinction is possible by examining the higher-order moments of an extended Fokker–Planck equation, i.e., the Kramers–Moyal equation, by comparing the scaling of moments for different finite differences. This, although hard to employ in timeseries analysis, is just a reflection of the Lindeberg continuity theorem [66], i.e., a process with jumps does not scale as the increments of a classical continuous Brownian equivalent, in a power relation with time (or space).

3.3 Software development

Results

In publications #3 and #7 two closely connected software programmes are developed. Publication #3 implements an elementary kernel-density estimator to extract Kramers–Moyal coefficients for stochastic processes in any dimension, based on a Nadaraya–Watson like estimator [121, 122] for stochastic processes [123, 124] It relies on a numerical convolution procedure to extract the coefficient in Fourier space, making it computationally efficient.

Publication #7 specialised the software in Publication #3 for jump-diffusion processes in one dimension, alongside the explicit derivation of a set of second-order corrections to the exponential representation of the Kramers–Moyal operator of jump-diffusion processes. The software extracts up to the eight order of the conditional moments up to eighth order of a (continuous-time) stochastic timeseries. As presented in Ref. [67] a relation between the parameters of a jump-diffusion process and the conditional moments up to eight order allows for a closed form solution, making it possible to invert the problem and retrieve the parameters non-parametrically from data, in a similar fashion to what is shown in Publication #3. The set of second-order corrections of the representation of the Kramers–Moyal operator puts the non-parametric recovery of the underlying parameters of the jump-diffusion process at equal footing to what was developed in Ref. [74].

Discussion

Non-parametric estimators are a common tool to uncover the parameters of stochastic timeseries, having been applied extensively for over two decades [78, 124, 66]. Their applicability relies on a set of criteria which is often hard to justify for real-world data: stationarity and Markovianity.

Firstly, the numerical method developed in Publication #3 results from a conventional usage of convolutional methods in Fourier space. This naturally offers considerable numerical speed-ups, but can find a natural drawback when examining jump-like events in stochastic processes, as these jumps are fundamentally discontinuous transitions in a timeseries, which in Fourier space can lie outside the Nyquist frequency in the numerical implementation of a discrete Fourier transform (this being the method employed in the Fourier transform).

Secondly, the developed second-order approximation of the Kramers–Moyal operator, relegated to the appendix in Publication #7 begs re-examination. Fundamentally the representation of the exponential representation of the operator, and the subsequent approximation of this exponential operator in a power series, leads to the emergence of terms with derivatives of the Kramers–Moyal coefficients. Naturally these are bounded in continuous diffusion, as the semi-group of the Brownian motion is bounded to second-order derivatives, but for discontinuous stochastic processes—such as the discussed jump-diffusion process—either an integral over a Lévy measure or an infinite sum of differential operators appears. The first-order approximation, i.e., the commonly employed estimators, does not see the emergence of derivative terms, but the second-order does, and so subsequently do the following orders. In the particular derivation for jump-diffusion processes presented in Publication #7 the terms with derivatives are discarded, under the assumption that the Kramers–Moyal coefficients are lower-order polynomials [74]. This assumption, absolutely fair, seems appropriate, but a verification of whether the derived moments are indeed correctly normalised when discarding these elements is yet to be provided. This truncation seems benign for diffusions, see Ref. [74], however the presence of infinite sums of Kramers–Moyal coefficients and their respective derivatives could become problematic for jump-diffusion processes.

Contemporaneous works

Numerical implementations of kernel density estimators are not uncommon in statistics software packages [125], but no known implementation of parametric or non-parametric

Kramers–Moyal coefficient estimators are known to exist in the computational language Python, but an implementation in the computational language R exists for one a two-dimensional timeseries [126]. Existing implementations are limited to work in real space and do not enjoy the computational speed-ups provided by been able to employ a convolution of the conditional moments with kernel in Fourier space. Implementing convolutional methods has the added flexibility of being generalisable to any dimension. Likewise, implementations specific to Poissonian jump-diffusion processes are too novel to have been thus far developed.

Bibliography

- [1] L. Rydin Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, and B. Schäfer. Data-Driven Model of the Power-Grid Frequency Dynamics. *IEEE Access* **8**, 2020, pp. 43082–43097. DOI: 10.1109/ACCESS.2020.2967834.
- [2] M. Anvari, L. Rydin Gorjão, M. Timme, D. Witthaut, B. Schäfer, and H. Kantz. Stochastic properties of the frequency dynamics in real and synthetic power grids. *Physical Review Research* **2**, 2020, p. 013339. DOI: 10.1103/PhysRevResearch.2.013339.
- [3] L. Rydin Gorjão and F. Meirinhos. `kramersmoyal`: Kramers–Moyal coefficients for stochastic processes. *Journal of Open Source Software* **4**(44), 2019, p. 1693. DOI: 10.21105/joss.01693.
- [4] L. Rydin Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, G. C. Yalcin, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer. Open database analysis of scaling and spatio-temporal properties of power grid frequencies. *Nature Communication* **11**, 2020, p. 6362. DOI: 10.1038/s41467-020-19732-7.
- [5] L. Rydin Gorjão, L. Vanfretti, D. Witthaut, C. Beck, and B. Schäfer. Phase and amplitude synchronisation in power-grid frequency fluctuations, 2020. *In preparation*.
- [6] L. Rydin Gorjão, J. Heysel, K. Lehnertz, and M. R. R. Tabar. Analysis and data-driven reconstruction of bivariate jump-diffusion processes. *Physical Review E* **100**, 2019, p. 062127. DOI: 10.1103/PhysRevE.100.062127.
- [7] L. Rydin Gorjão, D. Witthaut, and P. G. Lind. `JumpDiff`: A Python library for statistical inference of jump-diffusion processes in sets of measurements, 2020. submitted to the Journal of Statistical Software.

- [8] L. Rydin Gorjão, K. Riechers, F. Hassanibesheli, D. Witthaut, and P. G. Lind. Dansgaard–Oeschger events: Change in stability and jumps modelled via univariate and bivariate jump-diffusion processes, 2020. *In preparation*.
- [9] B. H. Obama. Presidential policy directive 21: Critical infrastructure security and resilience. *Washington, DC*, 2013. URL: <https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil>.
- [10] P. Kundur, N. J. Balu, and M. G. Lauby. *Power System Stability and Control*. 1st edition. Vol. 7. McGraw-Hill, New York, 1994. ISBN: 978-0-070-63515-9.
- [11] European Network of Transmission System Operators for Electricity (ENTSO-E). *Operation Handbook: P1 – Policy 1: Load-Frequency Control and Performance*. 2015. URL: https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/publications/entsoe/Operation_Handbook/Policy_1_final.pdf.
- [12] J. Machowski, J. Bialek, and J. Bumby. *Power System Dynamics: Stability and Control*. 2nd edition. John Wiley & Sons, Chichester, 2008. ISBN: 978-0-470-72558-0.
- [13] Y. Parag and B. K. Sovacool. Electricity market design for the prosumer era. *Nature Energy* **1**(4), 2016, p. 16032. DOI: 10.1038/nenergy.2016.32.
- [14] M. Rohden, A. Sorge, M. Timme, and D. Witthaut. Self-Organized Synchronization in Decentralized Power Grids. *Physical Review Letters* **109**, 6 2012, p. 064101. DOI: 10.1103/PhysRevLett.109.064101.
- [15] A. E. Motter, S. A. Myers, M. Anghel, and T. Nishikawa. Spontaneous synchrony in power-grid networks. *Nature Physics* **9**(3), 2013, pp. 191–197. DOI: 10.1038/nphys2535.
- [16] F. Dörfler, M. Chertkov, and F. Bullo. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences* **110**(6), 2013, pp. 2005–2010. DOI: 10.1073/pnas.1212134110.
- [17] L. Pagnier and P. Jacquod. Inertia location and slow network modes determine disturbance propagation in large-scale power grids. *PLOS ONE* **14**(3), 2019, pp. 1–17. DOI: 10.1371/journal.pone.0213550.

- [18] B. Hartmann, I. Vokony, and I. Táci. Effects of decreasing synchronous inertia on power system dynamics—Overview of recent experiences and marketisation of services. *International Transactions on Electrical Energy Systems* **29**(12), 2019, e12128. DOI: 10.1002/2050-7038.12128.
- [19] J. Fleer, S. Zurmühlen, J. Meyer, J. Badedá, P. Stenzel, J.-F. Hake, and D. U. Sauer. Techno-economic evaluation of battery energy storage systems on the primary control reserve market under consideration of price trends and bidding strategies. *Journal of Energy Storage* **17**, 2018, pp. 345–356. DOI: 10.1016/j.est.2018.03.008.
- [20] Y. Kuznetsov. *Elements of Applied Bifurcation Theory*. 3rd edition. Springer-Verlag, New York, 2004. ISBN: 978-0-387-21906-6. DOI: 10.1007/978-1-4757-3978-7.
- [21] G. Filatrella, N. F. Pedersen, and K. Wiesenfeld. Generalized coupling in the Kuramoto model. *Phys. Rev. E* **75**, 1 2007, p. 017201. DOI: 10.1103/PhysRevE.75.017201.
- [22] P. Vorobev, D. M. Greenwood, J. H. Bell, J. W. Bialek, P. C. Taylor, and K. Turitsyn. Deadbands, Droop, and Inertia Impact on Power System Frequency Distribution. *IEEE Transactions on Power Systems* **34**(4), 2019, pp. 3098–3108. DOI: 10.1109/TPWRS.2019.2895547.
- [23] C. Balestra, F. Kaiser, D. Manik, and D. Witthaut. Multistability in lossy power grids and oscillator networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **29**(12), 2019, p. 123119. DOI: 10.1063/1.5122739.
- [24] M. Klein, G. J. Rogers, and P. Kundur. A fundamental study of inter-area oscillations in power systems. *IEEE Transactions on power systems* **6**(3), 1991, pp. 914–921. DOI: 10.1109/59.119229.
- [25] G. Filatrella, A. H. Nielsen, and N. F. Pedersen. Analysis of a power grid using a Kuramoto-like model. *The European Physical Journal B* **61**(4), 2008, pp. 485–491. DOI: 10.1140/epjb/e2008-00098-8.
- [26] P. Pourbeik, P. S. Kundur, and C. W. Taylor. The anatomy of a power grid blackout - Root causes and dynamics of recent major blackouts. *IEEE Power and Energy Magazine* **4**(5), 2006, pp. 22–29. DOI: 10.1109/MPAE.2006.1687814.

- [27] Y. Yang, T. Nishikawa, and A. E. Motter. Small vulnerable sets determine large network cascades in power grids. *Science* **358**(6365), 2017. DOI: 10.1126/science.aan3184.
- [28] F. Kaiser, J. Strake, and D. Witthaut. Collective effects of link failures in linear flow networks. *New Journal of Physics* **22**(1), 2020, p. 013053. DOI: 10.1088/1367-2630/ab6793.
- [29] REN21 Members. *Renewables 2020–Global status reports*. REN21, 2020. ISBN: 978-3-948393-00-7. URL: <https://www.ren21.net/reports/global-status-report/>.
- [30] J. Meadowcroft. What about the politics? Sustainable development, transition management, and long term energy transitions. *Policy Sciences* **42**(4), 2009, p. 323. DOI: 10.1007/s11077-009-9097-z.
- [31] J. Rogelj, G. Luderer, R. C. Pietzcker, E. Kriegler, M. Schaeffer, V. Krey, and K. Riahi. Energy system transformations for limiting end-of-century warming to below 1.5 C. *Nature Climate Change* **5**(6), 2015, pp. 519–527. DOI: 10.1038/nclimate2572.
- [32] J. Markard. The next phase of the energy transition and its implications for research and policy. *Nature Energy* **3**(8), 2018, pp. 628–633. DOI: 10.1038/s41560-018-0171-7.
- [33] M. Victoria, K. Zhu, T. Brown, G. B. Andresen, and M. Greiner. Early decarbonisation of the European energy system pays off. *Nature Communications* **11**(1), 2020, p. 6223. DOI: 10.1038/s41467-020-20015-4.
- [34] K. Schmietendorf, J. Peinke, and O. Kamps. The impact of turbulent renewable energy production on power grid stability and quality. *The European Physical Journal B* **90**(11), 2017, p. 222. DOI: 10.1140/epjb/e2017-80352-8.
- [35] H. Hähne. Propagation of fluctuations and detection of hidden units in network dynamical systems. PhD thesis. 2019. URL: <https://oops.uni-oldenburg.de/id/eprint/4210>.
- [36] H. Holttinen. Impact of hourly wind power variations on the system operation in the Nordic countries. *Wind Energy* **8**(2), 2005, pp. 197–218. DOI: 10.1002/we.143.

- [37] J. Weber, M. Reyers, C. Beck, M. Timme, J. G. Pinto, D. Witthaut, and B. Schäfer. Wind Power Persistence Characterized by Superstatistics. *Scientific Reports* **9**(1), 2019, p. 19971. DOI: 10.1038/s41598-019-56286-1.
- [38] M. Anvari, G. Lohmann, M. Wächter, P. Milan, E. Lorenz, D. Heinemann, M. R. R. Tabar, and J. Peinke. Short term fluctuations of wind and solar power systems. *New Journal of Physics* **18**, 2016, p. 063027. DOI: 10.1088/1367-2630/18/6/063027.
- [39] R. Hesse, D. Turschner, and H.-P. Beck. Micro grid stabilization using the Virtual Synchronous Machine (VISMA). *Proceedings of the International Conference on Renewable Energies and Power Quality (ICREPQ'09), Valencia, Spain*. 2009, pp. 15–17. DOI: 10.24084/REPQJ07.472.
- [40] F. Milano, F. Dörfler, G. Hug, D. J. Hill, and G. Verbič. Foundations and Challenges of Low-Inertia Systems (Invited Paper). *2018 Power Systems Computation Conference (PSCC)*. 2018, pp. 1–25. DOI: 10.23919/PSCC.2018.8450880.
- [41] J. Schiffer, R. Ortega, A. Astolfi, J. Raisch, and T. Sezi. Conditions for stability of droop-controlled inverter-based microgrids. *Automatica* **50**(10), 2014, pp. 2457–2469. DOI: 10.1016/j.automatica.2014.08.009.
- [42] J. Schiffer, T. Seel, J. Raisch, and T. Sezi. Voltage Stability and Reactive Power Sharing in Inverter-Based Microgrids With Consensus-Based Distributed Voltage Control. *IEEE Transactions on Control Systems Technology* **24**(1), 2016, pp. 96–109. DOI: 10.1109/TCST.2015.2420622.
- [43] M. F. Wolff, K. Schmietendorf, P. G. Lind, O. Kamps, J. Peinke, and P. Maass. Heterogeneities in electricity grids strongly enhance non-Gaussian features of frequency fluctuations under stochastic power input. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **29**(10), 2019, p. 103149. DOI: 10.1063/1.5122986.
- [44] A. Ulbig, T. S. Borsche, and G. Andersson. Impact of low rotational inertia on power system stability and operation. *IFAC Proceedings Volumes* **47**(3), 2014, pp. 7290–7297. DOI: 10.3182/20140824-6-ZA-1003.02615.
- [45] F. Milano and R. Zárate-Miñano. A Systematic Method to Model Power Systems as Stochastic Differential Algebraic Equations. *IEEE Transactions on Power Systems* **28**(4), 2013, pp. 4537–4544. DOI: 10.1109/TPWRS.2013.2266441.

- [46] J. Hindes, P. Jacquod, and I. B. Schwartz. Network desynchronization by non-Gaussian fluctuations. *Physical Review E* **100**, 5 2019, p. 052314. DOI: 10.1103/PhysRevE.100.052314.
- [47] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar. Approaching complexity by stochastic methods: From biological systems to turbulence. *Physics Reports* **506**(5), 2011, pp. 87–162. DOI: 10.1016/j.physrep.2011.05.003.
- [48] T. Weißbach and E. Welfonder. High frequency deviations within the European power system: Origins and proposals for improvement. *2009 IEEE/PES Power Systems Conference and Exposition*. IEEE. 2009, pp. 1–6. DOI: 10.1109/PSCE.2009.4840180.
- [49] H. Risken. *The Fokker–Planck Equation*. 2nd edition. Springer, Berlin, 1984. ISBN: 978-3-540-61530-9. DOI: 10.1007/978-3-642-61544-3.
- [50] R. Bradley. *Paleoclimatology: Reconstructing Climates of the Quaternary*. 3rd edition. Academic Press, 2015. ISBN: 978-0-123-86913-5.
- [51] S. J. Johnsen, H. B. Clausen, W. Dansgaard, K. Fuhrer, N. Gundestrup, C. U. Hammer, P. Iversen, J. Jouzel, B. Stauffer, and J. P. Steffensen. Irregular glacial interstadials recorded in a new Greenland ice core. *Nature* **359**(6393), 1992, pp. 311–313. DOI: 10.1038/359311a0.
- [52] W. Dansgaard, S. J. Johnsen, H. B. Clausen, D. Dahl-Jensen, N. S. Gundestrup, C. U. Hammer, C. S. Hvidberg, J. P. Steffensen, A. E. Sveinbjörnsdottir, J. Jouzel, and G. Bond. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**(6434), 1993, pp. 218–220. DOI: 10.1038/364218a0.
- [53] S. O. Rasmussen, M. Bigler, S. P. Blockley, T. Blunier, S. L. Buchardt, H. B. Clausen, I. Cvijanovic, D. Dahl-Jensen, S. J. Johnsen, H. Fischer, V. Gkinis, M. Guillevic, W. Z. Hoek, J. J. Lowe, J. B. Pedro, T. Popp, I. K. Seierstad, J. P. Steffensen, A. M. Svensson, P. Vallelonga, B. M. Vinther, M. J. Walker, J. J. Wheatley, and M. Winstrup. A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records: refining and extending the INTIMATE event stratigraphy. *Quaternary Science Reviews* **106**, 2014, pp. 14–28. DOI: 10.1016/j.quascirev.2014.09.007.

- [54] U. Ruth, D. Wagenbach, J. P. Steffensen, and M. Bigler. Continuous record of microparticle concentration and size distribution in the central Greenland NGRIP ice core during the last glacial period. *Journal of Geophysical Research: Atmospheres* **108**(D3), 2003. DOI: 10.1029/2002JD002376.
- [55] K. K. Andersen, N. Azuma, J.-M. Barnola, M. Bigler, P. Biscaye, N. Caillon, J. Chappellaz, H. B. Clausen, D. Dahl-Jensen, H. Fischer, J. Flückiger, D. Fritzsche, Y. Fujii, K. Goto-Azuma, K. Grønvold, N. S. Gundestrup, M. Hansson, C. Huber, C. S. Hvidberg, S. J. Johnsen, U. Jonsell, J. Jouzel, S. Kipfstuhl, A. Landais, M. Leuenberger, R. Lorrain, V. Masson-Delmotte, H. Miller, H. Motoyama, H. Narita, T. Popp, S. O. Rasmussen, D. Raynaud, R. Rothlisberger, U. Ruth, D. Samyn, J. Schwander, H. Shoji, M.-L. Siggard-Andersen, J. P. Steffensen, T. Stocker, A. E. Sveinbjörnsdóttir, A. Svensson, M. Takata, J.-L. Tison, T. Thorsteinsson, O. Watanabe, F. Wilhelms, J. W. C. White, and N. G. I. C. P. members. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431**(7005), 2004, pp. 147–151. DOI: 10.1038/nature02805.
- [56] V. Gkinis, S. Simonsen, S. Buchardt, J. White, and B. Vinther. Water isotope diffusion rates from the NorthGRIP ice core for the last 16,000 years – Glaciological and paleoclimatic implications. *Earth and Planetary Science Letters* **405**, 2014, pp. 132–141. DOI: 10.1016/j.epsl.2014.08.022.
- [57] See https://en.wikipedia.org/wiki/Greenland_Ice_Sheet_Project and https://en.wikipedia.org/wiki/Greenland_ice_core_project, respectively, and references therein, respectively.
- [58] M. Schulz. On the 1470-year pacing of Dansgaard–Oeschger warm events. *Paleoceanography* **17**(2), 2002, pp. 4-1-4-9. DOI: 10.1029/2000PA000571.
- [59] P. D. Ditlevsen, M. S. Kristensen, and K. K. Andersen. The Recurrence Time of Dansgaard–Oeschger Events and Limits on the Possible Periodic Component. *Journal of Climate* **18**(14), 2005, pp. 2594–2603. DOI: 10.1175/JCLI3437.1.
- [60] P. D. Ditlevsen, K. K. Andersen, and A. Svensson. The DO-climate events are probably noise induced: statistical investigation of the claimed 1470 years cycle. *Climate of the Past* **3**(1), 2007, pp. 129–134. DOI: 10.5194/cp-3-129-2007.

- [61] J. Lynch-Stieglitz. The Atlantic Meridional Overturning Circulation and Abrupt Climate Change. *Annual Review of Marine Science* **9**(1), 2017, pp. 83–104. DOI: 10.1146/annurev-marine-010816-060415.
- [62] D. Kondrashov, S. Kravtsov, A. W. Robertson, and M. Ghil. A Hierarchy of Data-Based ENSO Models. *Journal of Climate* **18**(21), 2005, pp. 4425–4444. DOI: 10.1175/JCLI3567.1.
- [63] D. Kondrashov, M. D. Chekroun, and M. Ghil. Data-driven non-Markovian closure models. *Physica D: Nonlinear Phenomena* **297**, 2015, pp. 33–55. DOI: 10.1016/j.physd.2014.12.005.
- [64] P. D. Ditlevsen and O. D. Ditlevsen. On the Stochastic Nature of the Rapid Climate Shifts during the Last Ice Age. *Journal of Climate* **22**(2), 2009, pp. 446–457. DOI: 10.1175/2008JCLI2430.1.
- [65] Y. Li and J. Duan. A Data-Driven Approach for Discovering Stochastic Dynamical Systems with Non-Gaussian Lévy Noise. *preprint on ArXiv 2005.03769*, 2020. <https://arxiv.org/abs/2005.03769>.
- [66] M. R. R. Tabar. *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems*. 1st edition. Springer International Publishing, 2019. ISBN: 978-3-030-18471-1. DOI: 10.1007/978-3-030-18472-8.
- [67] M. Anvari, M. R. R. Tabar, J. Peinke, and K. Lehnertz. Disentangling the stochastic behavior of complex time series. *Scientific Reports* **6**, 2016, p. 35435. DOI: 10.1038/srep35435.
- [68] K. Lehnertz, L. Zabawa, and M. R. R. Tabar. Characterizing abrupt transitions in stochastic dynamics. *New Journal of Physics* **20**(11), 2018, p. 113043. DOI: 10.1088/1367-2630/aaf0d7.
- [69] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme. Non-Gaussian power grid frequency fluctuations characterized by Lévy-stable laws and superstatistics. *Nature Energy* **3**(2), 2018, pp. 119–126. DOI: 10.1038/s41560-017-0058-z.
- [70] K. Sharafutdinov, L. Rydin Gorjão, M. Matthiae, T. Faulwasser, and D. Witthaut. Rotor-angle versus voltage instability in the third-order model for synchronous generators. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28**(3), 2018, p. 033117. DOI: 10.1063/1.5002889.

- [71] L. Rydin Gorjão and D. Witthaut. *Dynamic stability of electric power grids: Tracking the interplay of the network structure, transmission losses and voltage dynamics*. <https://arxiv.org/abs/1908.10083>. 2019.
- [72] R. Jumar, H. Maass, B. Schäfer, L. Rydin Gorjão, and V. Hagenmeyer. Power grid frequency database. *preprint on ArXiv 2006.01771*, 2020. <https://arxiv.org/abs/2006.01771>.
- [73] L. Rydin Gorjão and G. Hassan. *Multifractal Detrended Fluctuation Analysis in Python*. Open-source software. 2020. DOI: 10.5281/zenodo.3996121.
- [74] J. Gottschall and J. Peinke. On the definition and handling of different drift and diffusion estimates. *New Journal of Physics* **10**, 2008, p. 083034. DOI: 10.1088/1367-2630/10/8/083034.
- [75] C. Beck and E. G. D. Cohen. Superstatistics. *Physica A* **322**, 2003, pp. 267–275. DOI: 10.1016/S0378-4371(03)00019-0.
- [76] C. Beck, E. G. D. Cohen, and H. L. Swinney. From time series to superstatistics. *Physical Review E* **72**(5), 2005, p. 056133. DOI: 10.1103/PhysRevE.72.056133.
- [77] S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, and Y. Dodge. Turbulent cascades in foreign exchange markets. *Nature* **381**(6585), 1996, pp. 767–770. DOI: 10.1038/381767a0.
- [78] R. Friedrich and J. Peinke. Description of a Turbulent Cascade by a Fokker-Planck Equation. *Physical Review Letters* **78**, 1997, pp. 863–866. DOI: 10.1103/PhysRevLett.78.863.
- [79] B. Castaing, Y. Gagne, and E. Hopfinger. Velocity probability density functions of high Reynolds number turbulence. *Physica D: Nonlinear Phenomena* **46**(2), 1990, pp. 177–200. DOI: 10.1016/0167-2789(90)90035-N.
- [80] B. Castaing. Scalar intermittency in the variational theory of turbulence. *Physica D: Nonlinear Phenomena* **73**(1), 1994, pp. 31–37. DOI: 10.1016/0167-2789(94)90223-2.
- [81] R. Metzler and J. Klafter. The random walk’s guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports* **339**(1), 2000, pp. 1–77. DOI: 10.1016/S0370-1573(00)00070-3.

- [82] R. Metzler and J. Klafter. The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. *Journal of Physics A: Mathematical and General* **37**(31), 2004, R161–R208. DOI: 10.1088/0305-4470/37/31/r01.
- [83] A. Arneodo, C. Baudet, F. Belin, R. Benzi, B. Castaing, B. Chabaud, R. Chavarria, S. Ciliberto, R. Camussi, F. Chillà, B. Dubrulle, Y. Gagne, B. Hebral, J. Herweijer, M. Marchand, J. Maurer, J. F. Muzy, A. Naert, A. Noullez, J. Peinke, F. Roux, P. Tabeling, W. van de Water, and H. Willaime. Structure functions in turbulence, in various flow configurations, at Reynolds number between 30 and 5000, using extended self-similarity. *Europhysics Letters (EPL)* **34**(6), 1996, pp. 411–416. DOI: 10.1209/epl/i1996-00472-2.
- [84] J. Corcuera, D. Nualart, and J. H. Woerner. Power variation of some integral fractional processes. *Bernoulli* **12**(4), 2006, pp. 713–735. DOI: 10.3150/bj/1155735933.
- [85] K. Kubilius, Y. Mishura, and K. Ralchenko. *Parameter estimation in fractional diffusion models*. 1st edition. Vol. 8. Springer, 2017. ISBN: 978-3-319-71029-7. DOI: 10.1007/978-3-319-71030-3.
- [86] J. Istas and G. Lang. Quadratic variations and estimation of the local Hölder index of a Gaussian process. *Annales de l’I.H.P. Probabilités et statistiques* **33**(4), 1997, pp. 407–436. URL: http://www.numdam.org/item/AIHPB_1997__33_4_407_0.
- [87] K. Kubilius and Y. Mishura. The rate of convergence of Hurst index estimate for the stochastic differential equation. *Stochastic Processes and their Applications* **122**(11), 2012, pp. 3718–3739. DOI: 10.1016/j.spa.2012.06.011.
- [88] Y. Hu, D. Nualart, and H. Zhou. Parameter estimation for fractional Ornstein–Uhlenbeck processes of general Hurst parameter. *Statistical Inference for Stochastic Processes* **22**(1), 2019, pp. 111–142. DOI: 10.1007/s11203-017-9168-2.
- [89] F. Comte and N. Marie. Nonparametric estimation in fractional SDE. *Statistical Inference for Stochastic Processes* **22**(3), 2019, pp. 359–382. DOI: 10.1007/s11203-019-09196-y.
- [90] J. A. Short, D. G. Infield, and L. L. Freris. Stabilization of Grid Frequency Through Dynamic Demand Control. *IEEE Transactions on Power Systems* **22**(3), 2007, pp. 1284–1293. DOI: 10.1109/TPWRS.2007.901489.

- [91] G. Weiss, F. Dörfler, and Y. Levron. A stability theorem for networks containing synchronous generators. *Systems & Control Letters* **134**, 2019, p. 104561. DOI: 10.1016/j.sysconle.2019.104561.
- [92] H. Hähne, J. Schottler, M. Waechter, J. Peinke, and O. Kamps. The footprint of atmospheric turbulence in power grid frequency measurements. *Europhysics Letters (EPL)* **121**(3), 2018, p. 30001. DOI: 10.1209/0295-5075/121/30001.
- [93] F. Dörfler and F. Bullo. Synchronization in complex networks of phase oscillators: A survey. *Automatica* **50**(6), 2014, pp. 1539–1564. DOI: 10.1016/j.automatica.2014.04.012.
- [94] M. Tyloo, L. Pagnier, and P. Jacquod. The key player problem in complex oscillator networks and electric power grids: Resistance centralities identify local vulnerabilities. *Science Advances* **5**(11), 2019. DOI: 10.1126/sciadv.aaw8359.
- [95] J. Strake, F. Kaiser, F. Basiri, H. Ronellenfitsch, and D. Witthaut. Non-local impact of link failures in linear flow networks. *New Journal of Physics* **21**(5), 2019, p. 053009. DOI: 10.1088/1367-2630/ab13ba.
- [96] X. Zhang, D. Witthaut, and M. Timme. Topological Determinants of Perturbation Spreading in Networks. *Physical Review Letters* **125**, 21 2020, p. 218301. DOI: 10.1103/PhysRevLett.125.218301.
- [97] B. Schäfer, D. Witthaut, M. Timme, and V. Latora. Dynamically induced cascading failures in power grids. *Nature Communications* **9**(1), 2018, p. 1975. DOI: 10.1038/s41467-018-04287-5.
- [98] H. Li, Y. Yuan, X. Zhang, and C. Su. Analysis of frequency emergency control characteristics of UHV AC/DC large receiving end power grid. *The Journal of Engineering* **2017**(13), 2017, pp. 686–690. DOI: 10.1049/joe.2017.0418.
- [99] K. Uhlen, S. Elenius, I. Norheim, J. Jyrinsalo, J. Elovaara, and E. Lakervi. Application of linear analysis for stability improvements in the Nordic power transmission system. *2003 IEEE Power Engineering Society General Meeting*. Vol. 4. 2003, pp. 2097–2103. DOI: 10.1109/PES.2003.1270938.
- [100] L. Vanfretti, R. García-Valle, K. Uhlen, E. Johansson, D. Trudnowski, J. W. Pierre, J. H. Chow, O. Samuelsson, J. Østergaard, and K. E. Martin. Estimation of Eastern Denmark’s electromechanical modes from ambient phasor measurement data. *2010 IEEE Power Engineering Society General Meeting*. 2010, pp. 1–8. DOI: 10.1109/PES.2010.5588149.

- [101] L. Vanfretti, L. Dosiek, J. W. Pierre, D. Trudnowski, J. H. Chow, R. García-Valle, and U. Aliyu. Application of ambient analysis techniques for the estimation of electromechanical oscillations from measured PMU data in four different power systems. *European Transactions on Electrical Power* **21**(4), 2011, pp. 1640–1656. DOI: 10.1002/etep.507.
- [102] K. Uhlen, L. Vanfretti, M. M. de Oliveira, A. B. Leirbukt, V. H. Aarstrand, and J. O. Gjerde. Wide-Area Power Oscillation Damper implementation and testing in the Norwegian transmission network. *2012 IEEE Power and Energy Society General Meeting*. 2012, pp. 1–7. DOI: 10.1109/PESGM.2012.6344837.
- [103] H. Hähne, K. Schmietendorf, S. Tamrakar, J. Peinke, and S. Kettemann. Propagation of wind-power-induced fluctuations in power grids. *Physical Review E* **99**, 5 2019, p. 050301. DOI: 10.1103/PhysRevE.99.050301.
- [104] P. G. Meyer, M. Anvari, and H. Kantz. Identifying characteristic time scales in power grid frequency fluctuations with DFA. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **30**(1), 2020, p. 013130. DOI: 10.1063/1.5123778.
- [105] S. Pfenninger, J. DeCarolus, L. Hirth, S. Quoilin, and I. Staffell. The importance of open data and software: Is energy research lagging behind? *Energy Policy* **101**(C), 2017, pp. 211–215. DOI: 10.1016/j.enpol.2016.11.046.
- [106] S. Pfenninger, L. Hirth, I. Schlecht, E. Schmid, F. Wiese, T. Brown, C. Davis, M. Gidden, H. Heinrichs, C. Heuberger, S. Hilpert, U. Krien, C. Matke, A. Nebel, R. Morrison, B. Müller, G. Pleßmann, M. Reeg, J. C. Richstein, A. Shivakumar, I. Staffell, T. Tröndle, and C. Wingenbach. Opening the black box of energy modelling: Strategies and lessons learned. *Energy Strategy Reviews* **19**, 2018, pp. 63–71. DOI: 10.1016/j.esr.2017.12.002.
- [107] T. Braun, M. Wächter, J. Peinke, and T. Guhr. Correlated power time series of individual wind turbines: A data driven model approach. *Journal of Renewable and Sustainable Energy* **12**(2), 2020, p. 023301. DOI: 10.1063/1.5139039.
- [108] J. C. Cox, J. E. Ingersoll, and S. A. Ross. A Theory of the Term Structure of Interest Rates. *Econometrica* **53**(2), 1985, pp. 385–407. DOI: 10.2307/1911242.
- [109] H. Geman and Y. F. Shih. Modeling Commodity Prices under the CEV Model. *The Journal of Alternative Investments* **11**(3), 2008, pp. 65–84. DOI: 10.3905/JAI.2009.11.3.065.

- [110] J. L. Forman and M. Sørensen. The Pearson Diffusions: A Class of Statistically Tractable Diffusion Processes. *Scandinavian Journal of Statistics* **35**(3), 2008, pp. 438–465. URL: <https://www.jstor.org/stable/41000281>.
- [111] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E* **49**, 1994, pp. 1685–1689. DOI: 10.1103/PhysRevE.49.1685.
- [112] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A* **316**(1), 2002, pp. 87–114. DOI: 10.1016/S0378-4371(02)01383-3.
- [113] N. Boers, M. D. Chekroun, H. Liu, D. Kondrashov, D.-D. Rousseau, A. Svensson, M. Bigler, and M. Ghil. Inverse stochastic–dynamic models for high-resolution Greenland ice core records. *Earth System Dynamics* **8**(4), 2017, pp. 1171–1190. DOI: 10.5194/esd-8-1171-2017.
- [114] J. Muzy, E. Bacry, and A. Arneodo. The multifractal formalism revisited with wavelets. *International Journal of Bifurcation and Chaos* **04**(02), 1994, pp. 245–302. DOI: 10.1142/S0218127494000204.
- [115] B. Øksendal and A. Sulem. *Applied Stochastic Control of Jump Diffusions*. 1st edition. Springer, Cham, 2019. ISBN: 978-3-030-02781-0. DOI: 10.1007/978-3-030-02781-0.
- [116] D. R. Baños, F. Cordoni, G. di Nunno, L. di Persio, and E. E. Røse. Stochastic systems with memory and jumps. *Journal of Differential Equations* **266**(9), 2019, pp. 5772–5820. DOI: 10.1016/j.jde.2018.10.052.
- [117] R. Huisman and R. Mahieu. Regime jumps in electricity prices. *Energy Economics* **25**(5), 2003, pp. 425–434. DOI: 10.1016/S0140-9883(03)00041-0.
- [118] F. E. Benth, J. Š. Benth, and S. Koekebakker. *Stochastic Modeling of Electricity and Related Markets*. 1st edition. World Scientific, 2008. ISBN: 978-981-281-230-8. DOI: 10.1142/6811.
- [119] P. Tankov and R. Cont. *Financial Modelling with Jump Processes*. 1st edition. Chapman and Hall/CRC, 2003. ISBN: 978-1-584-88413-2.
- [120] D. Applebaum. *Lévy Processes and Stochastic Calculus*. 2nd edition. Cambridge University Press, Cambridge, 2009. ISBN: 978-0-521-73865-1. DOI: 10.1017/CB09780511809781.

- [121] E. A. Nadaraya. On Estimating Regression. *Theory of Probability & Its Applications* **9**(1), 1964, pp. 141–142. DOI: 10.1137/1109020.
- [122] G. S. Watson. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A* **26**(4), 1964, pp. 359–372. ISSN: 0581572X. URL: <http://www.jstor.org/stable/25049340>.
- [123] J. Prusseit and K. Lehnertz. Stochastic Qualifiers of Epileptic Brain Dynamics. *Physical Review Lett.* **98**, 2007, p. 138103. DOI: 10.1103/PhysRevLett.98.138103.
- [124] D. Lamouroux and K. Lehnertz. Kernel-based regression of drift and diffusion coefficients of stochastic processes. *Physics Letters A* **373**(39), 2009, pp. 3507–3512. DOI: 10.1016/j.physleta.2009.07.073.
- [125] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2011, pp. 2825–2830. URL: <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [126] P. Rinn, P. G. Lind, M. Wächter, and J. Peinke. The Langevin Approach: An R Package for Modeling Markov Processes. *Journal of Open Research Software* **4**(1), 2016, e34. DOI: 10.5334/jors.123.
- [127] Réseau de Transport d’Électricité (RTE). *Network frequency*. https://clients.rte-france.com/lang/an/visiteurs/vie/vie_frequence.jsp.
- [128] TransnetBW GmbH. *Regelenergie Bedarf + Abruf*. <https://www.transnetbw.de/de/strommarkt/systemdienstleistungen/regelenergie-bedarf-und-abruf>.
- [129] Fingrid. *Nordic power system frequency measurement data*. <https://data.fingrid.fi/en/dataset/frequency-historical-data>.
- [130] National Grid ESO. *Historic frequency data*. <https://www.nationalgrideso.com/balancing-services/frequency-response-services/historic-frequency-data>.
- [131] R. Jumar, H. Maass, B. Schäfer, L. Rydin Gorjão, and V. Hagenmeyer. *Power grid frequency data base*. <https://osf.io/by5hu>. 2020.

- [132] J. P. Steffensen, Centre for Ice and Climate, Niels Bohr Institute. *Data, icesamples and software*. 2014. URL: <https://www.iceandclimate.nbi.ku.dk/data/>.
- [133] I. K. Seierstad, P. M. Abbott, M. Bigler, T. Blunier, A. J. Bourne, E. Brook, S. L. Buchardt, C. Buizert, H. B. Clausen, E. Cook, D. Dahl-Jensen, S. M. Davies, M. Guillevic, S. J. Johnsen, D. S. Pedersen, T. J. Popp, S. O. Rasmussen, J. P. Severinghaus, A. Svensson, and B. M. Vinther. Consistently dated records from the Greenland GRIP, GISP2 and NGRIP ice cores for the past 104 ka reveal regional millennial-scale $\delta^{18}O$ gradients with possible Heinrich event imprint. *Quaternary Science Reviews* **106**, 2014, pp. 29–46. DOI: 10.1016/j.quascirev.2014.10.032.

Appendix A

Author contributions

Publication #1

L. Rydin Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, and B. Schäfer. *Data-Driven Model of the Power-Grid Frequency Dynamics*. IEEE Access **8**, 2020, pp. 43082–43097, Ref. [1].

Author contributions The model, its implementation, all data collection, and analysis is the work of the first author. The particularity of the model choice was discussed with all authors. Discussion and conclusions drawn in the paper are a common effort.

Publication #2

M. Anvari, L. Rydin Gorjão, M. Timme, D. Witthaut, B. Schäfer, and H. Kantz. *Stochastic properties of the frequency dynamics in real and synthetic power grids*. Physical Review Research **2**(1), 2020, p. 013339. Ref. [2].

Author contributions The statistical properties of extensive physics drawn from the data are the work of the first author. The synthetic model is work of the second author. The non-extensive physical properties are authored by the penultimate author. The particularity of the statistical interpretation was discussed with all authors. Discussion and conclusions drawn in the paper are a common effort.

Publication #4

L. Rydin Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer. *Open data base analysis of scaling and spatio-temporal properties of power grid frequencies*. Nature Communications **11**, p. 6362, 2020, Ref. [4].

Author contributions The analysis is the work of the first-author and last author. The data collection was work of the second and third author. The interpretation of the Karhunen–Loève results are a joint effort of the first, fifth and last author. Discussion and conclusions drawn in the paper are a common effort.

Publication #5

L. Rydin Gorjão, L. Vanfretti, D. Witthaut, C. Beck and B. Schäfer, under the working title *Phase and amplitude synchronisation in power-grid frequency fluctuations*, Ref. [5].

Author contributions The analysis is the work of the first-author. The provision of the data is work of the second author. Discussion and conclusions drawn in the paper are a common effort.

Publication #3

L. Rydin Gorjão and F. Meirinhos. `kramersmoyal`: Kramers–Moyal coefficients for stochastic processes. Journal of Open Source Software **4**(44), 2019, p. 1693, Ref. [3].

Author contributions The theoretical background is work of the first author. The prototypical numerical procedure in one and two dimensions is work of the first author, generalised to N dimension by the second author. The discussion is work of the first author.

Publication #6

L. Rydin Gorjão, J. Heysel, K. Lehnertz, and M. R. R. Tabar. *Analysis and data-driven reconstruction of bivariate jump-diffusion processes*. Physical Review E **100**, 2019, p. 062127, Ref. [6].

Author contributions The theoretical background is work of the two last authors. The implementation, analysis, and employed study models is work of the first author. Discussion and conclusions drawn in the paper are a common effort.

Publication #7

L. Rydin Gorjão, D. Witthaut, and P. G. Lind. *JumpDiff: A Python library for statistical inference of jump-diffusion processes in sets of measurements*, submitted to the Journal of Statistical Software, 2020, Ref. [7].

Author contributions The theoretical background is joint work of the first and last author. The implementation, analysis, and numerical development is work of the first author. Discussion and conclusions drawn in the paper are a common effort.

Publication #8

L. Rydin Gorjão, K. Riechers, F. Hassanibesheli, D. Witthaut, and P. G. Lind, under the working title *Dansgaard–Oeschger events: Change in stability and jumps modelled via univariate and bivariate jump-diffusion processes*, Ref. [8].

Author contributions The model implementation, analysis, and conclusions are work of the first author. The field-specific interpretation is work of the first three authors. Discussion and conclusions drawn in the paper are a common effort.

Appendix B

Data availability

B.1 Power-grid systems

Power-grid frequency is more familiarly known as the 50 Hz in Europe, Russian, India, China, and parts of Japan (60 Hz in the United States of America and other parts of Japan). Power-grid frequency data is rarely freely available across different power-grid systems in the world. In Continental Europe, which comprises the majority of countries in Europe and a large array of different power-grid operators, only two recordings of power-grid frequency are freely available, provided by the French *Réseau de Transport d'Électricité (RTE)* [127] and the German *TransnetBW* [128]. Likewise in the Nordic grid, comprising Norway, Sweden, Finland, and a small part of Denmark, the only available source of data is provided by the Finnish operating system *FinGrid* [129]. Lastly, the operating system in Great Britain is controlled by the *National Grid ESO* [130], providing also a single recording at a single location.

A portion of this thesis comprised a concerted effort in acquiring more power-grid frequency data, which is openly available in Ref. [131], and is documented in Ref. [72].

Lastly, a set of synchronous recordings from power-grid frequency measurements from the Nordic synchronous areas were utilised in Publication #5. Unfortunately, this data set has been provided under Non-Disclosure Agreement and thus cannot be made freely available. Hopefully the Non-Disclosure Agreement will be lifted in the future such that the data set can be published along with the article [5].

B.1.1 List of sources

Links as of the 21st of December 2020.

Publication #1: Continental European data [128]. Great Britain [130].

Publication #2: Continental European data [127, 128]. Great Britain [130].

Publication #4: All data [131, 72]. Documented in lrydin.github.io/Power-Grid-Frequency/.

Publication #5: Protected under Non-Disclosure Agreement.

B.2 Paleo-climate high-frequency data

The data employed in this thesis relates stems from the *North Greenland Ice Core Project* NGRIP project [54, 55, 56]. The data are freely available with the aforementioned publications, with the care for particular pre-processing employed by the authors of each work. Moreover, for comparison, the data from the *Greenland Ice Sheet Project Two* (GISP2) and the *Greenland Ice Core Project* (GRIP) were employed [57]. The analysis is solely conveyed on the data from the NGRIP project.

B.2.1 List of sources

Links as of the 21st of December 2020.

Publication #8: NGRIP project data www.iceandclimate.nbi.ku.dk/data/. Documented in [132, 133].

Appendix C

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie—abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten—noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

- #1 L. Rydin Gorjão, M. Anvari, H. Kantz, C. Beck, D. Witthaut, M. Timme, and B. Schäfer. *Data-driven model of the power-grid frequency dynamics*, IEEE Access **8**, pp. 43082–43097 (2020).
doi:10.1109/ACCESS.2020.2967834.

- #2 M. Anvari, L. Rydin Gorjão, M. Timme, D. Witthaut, B. Schäfer, and H. Kantz., *Stochastic properties of the frequency dynamics in real and synthetic power grids*, Physical Review Research **2**, p. 013339 (2020).
doi:10.1103/PhysRevResearch.2.013339.
- #4 L. Rydin Gorjão, R. Jumar, H. Maass, V. Hagenmeyer, G. C. Yalcin, J. Kruse, M. Timme, C. Beck, D. Witthaut, and B. Schäfer. *Open database analysis of scaling and spatio-temporal properties of power grid frequencies*, Nature Communications **11**, p. 6362 (2020).
doi:10.1038/s41467-020-19732-7.
- #5 L. Rydin Gorjão and F. Meirinhos. *kramersmoyal: Kramers–Moyal coefficients for stochastic processes*, Journal of Open Source Software **4**(44), p. 1693 (2019).
doi:10.21105/joss.01693.
- #3 L. Rydin Gorjão, J. Heysel, K. Lehnertz, and M. R. R. Tabar. *Analysis and data-driven reconstruction of bivariate jump-diffusion processes*, Physical Review E **100**, p. 062127 (2019).
doi:10.1103/PhysRevE.100.062127.

Ein weiteres Manuskript wurde zur Veröffentlichung bei einem Journal eingereicht, zwei weitere sind noch in Arbeit.



21.12.2020, Leonardo Rydin Gorjão

